DRAFT --- DRAFT --- DRAFT

NSF'S CYBERINFRASTRUCTURE VISION FOR 21ST CENTURY DISCOVERY

NSF Cyberinfrastructure Council

National Science Foundation
July 20, 2006
Version 7.1

ACRONYMS

CCSDS	Consultative Committee for Space Data Standards
CI	Cyberinfrastructure
CIO	Chief Information Officer
CODATA	National Committee on Data for Science and Technology
CPU	Central Processing Unit
CSNET	Computer Science Network
DARPA	Defense Advanced Research Projects Agency
DOD	Department of Defense
DOE	Department of Energy
ETF	Extensible Terascale Facility
FLOPS	Floating point operations/sec
HPC	High Performance Computing
HPCMOD	DOD's High-Performance Computing Modernization program
HPCS	DARPA's High Productivity Computing Systems program
HPCC	High-Performance Computing and Communications
GEON	Geosciences Network
GriPhyN	Grid Physics Network
ICPSR	Inter-university Consortium for Political and Social Research
ICSU	International Council for Science
ICSTI	International Council for Scientific and Technical Information
IRIS	Incorporated Research Institutions for Seismology
ISO	International Organization of Standardization
IT	Information Technology
ITR	Information Technology Research
IVDGL	International Virtual Data Grid Laboratory
MPI	Message Passing Interface
NARA	National Archives and Record Administration
NASA	National Aeronautics and Space Administration
NIH	National Institutes of Health
NITRD	Networking and Information Technology Research and Development
NNSA	National Nuclear Security Administration
NSFNET	NSF Network
NRC	National Research Council
NSB	National Science Board
NSF	National Science Foundation
OAIS	Open Archival Information System
OS	Operating System
PACI	Partnership for Advanced Computational Infrastructure
PITAC	President's Information Technology Advisory Committee
RLG	Research Library Group
SciDAC	Scientific Discovery through Advanced Computing
SSP	Software Services Provider
TFLOPS	Teraflops: Trillion floating point operations/sec
USNC/ CODATA	U.S. National Committee for CODATA
VO	Virtual Organization
WDC	World Data Center

Acronyms

TABLE OF CONTENTS

CHAPTER 1
CALL TO ACTION

I. DRIVERS AND OPPORTUNITIES

How does a protein fold? What happens to space-time when two black holes collide? What impact does species gene flow have on an ecological community? What are the key factors that drive climate change? Did one of the trillions of collisions at the Large Hadron Collider produce a Higgs boson, the dark matter particle or a black hole? Can we create an individualized model of each human being for personalized healthcare delivery? How does major technological change affect human behavior and structure complex social relationships? What answers will we find – to questions we have yet to ask – in the very large datasets that are being produced by telescopes, sensor networks, and other experimental facilities?

These questions – and many others – are only now coming within our ability to answer because of advances in computing and related information technology. Once used by a handful of elite researchers in a few research communities on select problems, advanced computing has become essential to future progress across the frontier of science and engineering. Coupled with continuing improvements in microprocessor speeds, converging advances in networking, software, visualization, data systems and collaboration platforms are changing the way research and education is accomplished.

Today's scientists and engineers need access to new information technology capabilities, such as distributed wired and wireless observing network complexes, and sophisticated simulation tools that permit exploration of phenomena that can never be observed or replicated by experiment. Computation offers new models of behavior and modes of scientific discovery that greatly extend the limited range of models that can be produced with mathematics alone, for example, chaotic behavior. Fewer and fewer researchers working at the frontiers of knowledge can carry out their work without cyberinfrastructure of one form or another.

While hardware performance has been growing exponentially – with gate density doubling every 18 months, storage capacity every 12 months, and network capability every 9 months – it has become clear that increasingly capable hardware is not the only requirement for computation-enabled discovery. Sophisticated software, visualization tools, middleware and scientific applications created and used by interdisciplinary teams are critical to turning flops, bytes and bits into scientific breakthroughs. The exploration of new organizational models and the creation of enabling policies and processes are also essential. It is the combined power of these capabilities and approaches that is necessary to advance the frontiers of science and engineering, to make seemingly intractable problems solvable and to pose profound new scientific questions.

The comprehensive infrastructure needed to capitalize on dramatic advances in information technology has been termed cyberinfrastructure. Cyberinfrastructure integrates hardware for computing, data and networks, digitally-enabled sensors, observatories and experimental facilities, and an interoperable suite of software and middleware services and tools. Investments in interdisciplinary teams and cyberinfrastructure professionals with expertise in algorithm development, system operations, and applications development are also essential to exploit the full power of cyberinfrastructure to create, disseminate, and preserve scientific data, information, and knowledge.

For four decades, NSF has provided leadership in the scientific revolution made possible by information technology (Appendices A and B). Through investments ranging from supercomputing centers and the Internet to software and algorithm development, information technology has stimulated scientific breakthroughs across all science and engineering fields. Most recently, NSF's Information Technology Research (ITR) priority area sowed the seeds of broad and intensive collaboration between the computational, computer and domain research communities that sets the stage for this "Call to Action."

NSF is the only agency within the U.S. government that funds research and education across all disciplines of science and engineering. Over the past five years, NSF has held community workshops, commissioned blue-ribbon panels and carried out extensive internal planning (Appendix A.) Thus, it is strategically placed to leverage, coordinate and transition cyberinfrastructure advances in one field to all fields of research.

Other Federal agencies, the Administration and Congress, the private sector, and other nations are aware of the growing importance of cyberinfrastructure to progress in science and engineering. Other Federal agencies have planned improved capabilities for specific disciplines, and in some cases to address interdisciplinary challenges. Other countries have also been making significant progress in scientific cyberinfrastructure. Thus, the U.S. must engage in and actively benefit from cyberinfrastructure developments around the world.

Not only is the time ripe for a coordinated investment in cyberinfrastructure, progress at the science and engineering frontiers depends upon it. Our communities are in place and are poised to respond to such an investment.

Working with the science and engineering research and education communities and partnering with other key stakeholders, NSF is ready to lead.

II. VISION, MISSION, AND PRINCIPLES

A. Vision

NSF will play a leadership role in the development and support of a comprehensive cyberinfrastructure essential to 21st century advances in science and engineering research and education.

B. Mission

NSF's mission for cyberinfrastructure (CI) is to:

- Develop a human-centered CI that is driven by science and engineering research and education opportunities;
- Provide the science and engineering communities with access to world-class CI tools and services, including those focused on: high performance computing; data, data analysis and visualization; cyber-services and virtual organizations; and, learning and workforce development;
- Promote a CI that serves as an agent for broadening participation and strengthening the Nation's workforce in all areas of science and engineering;
- Provide a sustainable CI that is secure, efficient, reliable, accessible, usable, and interoperable, and which evolves as an essential national infrastructure for conducting science and engineering research and education; and
- Create a stable CI environment that enables the research and education communities to contribute to the agency's statutory mission.

C. Principles

The following principles will guide NSF's actions.

- Science and engineering research and education are foundational drivers of CI.
- NSF has a unique leadership role in formulating and implementing a national CI agenda focused on advancing science and engineering.
- Inclusive strategic planning is required to effectively address CI needs across a broad spectrum of organizations, institutions, communities and individuals, with input to the process provided through public comments, workshops, funded studies, advisory committees, merit review and open competitions.
- Strategic investments in CI resources and services are essential to continued U.S. leadership in science and engineering.
- The integration and sharing of cyberinfrastructure assets deployed and supported at national, regional, local, community, and campus levels represent the most effective way of constructing a comprehensive CI ecosystem suited to meeting future needs.
- National and international partnerships, public and private, that integrate CI users and providers and benefit NSF's research and education communities are also essential for enabling next-generation science and engineering.
- Existing strengths, including research programs and CI facilities, serve as a foundation upon which to build a CI designed to meet the needs of the broad science and engineering community.
- Merit review is essential for ensuring that the best ideas are pursued in all areas of CI funding.
- Regular evaluation and assessment tailored to individual projects is essential for ensuring accountability to all stakeholders.
- A collaborative CI governance structure that includes representatives who contribute to basic CI research, development and deployment, as well as those who use CI, is

essential to ensure that CI is responsive to community needs and empowers research at the frontier.

III. GOALS AND STRATEGIES

NSF's vision and mission statements need well-defined goals and strategies to turn them into reality. The goals underlying these statements are provided below, with each goal followed by a brief description of the strategy to achieve the goal.

Across the CI landscape, NSF will:

- ***Provide communities addressing the most computationally challenging problems with access to a world-class high performance computing (HPC) environment through NSF acquisition and through exchange-of-service agreements with other entities, where possible.***

NSF's investment strategy in the provision of HPC resources and services will be linked to careful requirements analyses of the computational needs of research and education communities. Our investments will be coordinated with those of other agencies in order to maximize access to these capabilities and to provide a range of representative high performance architectures.

- ***Broaden access to state-of-the-art computing resources, focusing especially on institutions with less capability and communities where computational science is an emerging activity.***

Building on the achievements of current CI service providers and other NSF investments, the agency will work to make necessary computing resources more broadly available, paying particular attention to emerging and underserved communities.

- ***Support the development and maintenance of robust systems software, programming tools, and applications needed to close the growing gap between peak performance and sustained performance on actual research codes, and to make the use of HPC systems, as well as novel architectures, easier and more accessible.***

NSF will build on research in computer science and other research areas to provide science and engineering applications and problem-solving environments that more effectively exploit innovative architectures and large-scale computing systems. NSF will continue and build upon its existing collaborations with other agencies in support of the development of HPC software and tools.

- ***Support the continued development, expansion, hardening and maintenance of end-to-end software systems – user interfaces, workflow engines, science and engineering applications, data management, analysis and visualization tools, collaborative tools, and other software integrated ito complete science and engineering systems via middleware – to bring the full power of a national cyberinfratructure to communities of scientists and engineers.***

These investments will build on the software products of current and former programs, and will leverage work in core computer science research and development efforts supported by NSF and other federal agencies.

- ***Support the development of the computing professionals, interdisciplinary teams, enabling policies and procedures, and new organizational structures such as virtual organizations, needed to achieve the scientific breakthroughs made possible by advanced CI, paying particular attention to the opportunities to broaden the participation of underrepresented groups.***

NSF will continue to invest in understanding how participants in its research and education communities, as well as the scientific workforce, can use CI. For example, virtual organizations empower communities of users to interact, exchange information and access and share resources through tailored interfaces. Some of NSF's investments will focus on appropriate mechanisms or structures for use, while others will focus on how best to train future users of CI. NSF will take advantage of the emerging communities associated with CI that provide unique and special opportunities for broadening participation in the science and engineering enterprise.

- ***Support state-of-the-art innovation in data management and distribution systems, including digital libraries and educational environments that are expected to contribute to many of the scientific breakthroughs of the 21st century.***

NSF will foster communication between forefront data management and distribution systems, digital libraries and other education environments sponsored in its various directorates. NSF will ensure that its efforts take advantage of innovation in large data management and distribution activities sponsored by other agencies and international efforts as well. These developments will play a critical role in decisions that NSF makes about long-lived data.

- ***Support the design and development of the CI needed to realize the full scientific potential of NSF's investments in tools and large facilities, from observatories and accelerators to sensor networks and remote observing systems.***

NSF's large facilities and other TOOLS investments require new types of CI such as wireless control of networks of sensors in hostile environments, rapid distribution and analysis of petascale data sets around the world, adaptive knowledge-based control and sampling systems, and innovative visualization systems for collaboration. NSF will ensure that these projects invest appropriately in CI capabilities, promoting the integrated and widespread use of the unique services provided by these and other facilities. In addition, NSF's CI programs will be designed to serve the needs of these projects.

- ***Support the development and maintenance of the increasingly sophisticated applications needed to achieve the scientific goals of research and education communities.***

The applications needed to produce cutting-edge science and engineering have become increasingly complex. They require teams, even communities, to develop and sustain wide and long-term applicability, and they leverage underlying software tool and increasingly common, persistent CI resources such as data repositories and authentication and authorization services. NSF's investments in applications will involve its directorates, which support domain-

specific science and engineering. Special attention will be paid to the cross-disciplinary nature of much of the work.

- ***Invest in the high-risk/high-gain basic research in computer science, computing and storage devices, mathematical algorithms and the human/CI interfaces that are critical to powering the future exponential growth in all aspects of computing, from hardware speed, storage, connectivity and scientific productivity.***

NSF's investments in operational CI must be coupled with vigorous research programs in the directorates that will ensure operational capabilities continue to expand and extend in the future. Important among these are activities to understand how humans adopt and use CI. NSF is especially placed to foster collaborations among computer scientists, social, behavioral and economic scientists, and other domain scientists and engineers to understand how humans can best use CI, both in research and education environments.

- ***Provide a framework that will sustain reliable, stable resources and services while enabling the integration of new technologies and research developments with a minimum of disruption to users.***

NSF will minimize disruption to users by realizing a comprehensive CI with an architecture and framework that emphasizes interoperability and open standards, providing flexibility for upgrades, enhancements and evolutionary changes. Pre-planned arrangements for alternative CI availabilities during competitions, changeovers and upgrades to production operations and services will be made, including cooperative arrangements with other agencies.

A strategy common to achieving all of these goals is partnering nationally and internationally, with other agencies, the private sector, and with universities to achieve a worldwide CI that is interoperable, flexible, efficient, evolving and broadly accessible. In particular, NSF will take a lead role in formulating and implementing a national CI strategy.

IV. PLANNING FOR CYBERINFRASTRUCTURE

To implement its cyberinfrastructure vision, NSF will develop interdependent plans for each of the following aspects of CI, with emphasis on their integration to create a balanced, science- and engineering-driven national CI:

- High Performance Computing;
- Data, Data Analysis, and Visualization;
- Cyber-services and Virtual Organizations; and
- Learning and Workforce Development.

Others may be added at a later date.

These plans will be reviewed annually and will evolve over time, paced by the considerable rate of innovation in computing and the growing needs of the science and engineering community for state-of-the-art CI capabilities. Through their simultaneous implementation, NSF's vision will become reality.

++

CHAPTER 2
PLAN FOR HIGH PERFORMANCE COMPUTING (2006-2010)

I. WHAT DOES HIGH PERFORMANCE COMPUTING OFFER SCIENCE AND ENGINEERING?

What are the three-dimensional structures of all of the proteins encoded by the human genome and how does structure influence their function in a human cell? What patterns of emergent behavior occur in models of very large societies? How do massive stars explode and produce the heaviest elements in the periodic table? What sort of abrupt transitions can occur in Earth's climate and ecosystem structure? How do these occur and under what circumstances? If we could design catalysts atom-by-atom, could we transform industrial synthesis? What strategies might be developed to optimize management of complex infrastructure systems? What kind of language processing can occur in large assemblages of neurons? Can we enable integrated planning and response to natural and man-made disasters that prevent or minimize the loss of life and property? These are just some of the important questions that researchers wish to answer using contemporary tools in a state-of-the-art High Performance Computing (HPC) environment.

With HPC tools, researchers study the properties of minerals at the extreme temperatures and pressures that occur deep within the Earth. They simulate the development of structure in the early Universe. They probe the structure of novel phases of matter such as the quark-gluon plasma. HPC capabilities enable the modeling of life cycles that capture interdependencies across diverse disciplines and multiple scales to create globally competitive manufacturing enterprise systems. And they examine the way proteins fold and vibrate after they are synthesized inside an organism. In fact, sophisticated numerical simulations permit scientists and engineers to perform a wide range of *in silico* experiments that would otherwise be too difficult, too expensive or impossible to perform in the laboratory.

HPC systems and services are also essential to the success of research conducted with sophisticated experimental tools. For example, without the waveforms produced by numerical simulation of black hole collisions and other astrophysical events, gravitational wave signals cannot be extracted from the data produced by the Laser Interferometer Gravitational Wave Observatory; high-resolution seismic inversions from the higher density of broad-band seismic observations furnished by the Earthscope project are necessary to determine shallow and deep Earth structure; simultaneous integrated computational and experimental testing is conducted on the Network for Earthquake Engineering Simulation to improve seismic design of buildings and bridges; and HPC is essential to extracting the signature of the Higgs boson and supersymmetric particles – two of the scientific drivers of the Large Hadron Collider – from the petabytes of data produced in the trillions of particle collisions.

Science and engineering research and education enabled by state-of-the-art HPC tools have a direct bearing on the Nation's competitiveness. If investments in HPC are to have long-term

impact on problems of national need, such as bioengineering, critical infrastructure protection (for example, the electric power grid), health care, manufacturing, nanotechnology, energy, and transportation, then HPC tools must deliver high performance capability to a wide range of science and engineering applications.

II. THE NEXT FIVE YEARS: CREATING A HIGH PERFORMANCE COMPUTING ENVIRONMENT FOR PETASCALE SCIENCE AND ENGINEERING

NSF's five-year HPC goal is to enable *petascale* science and engineering through the deployment and support of a world-class HPC environment comprising the most capable combination of HPC assets available to the academic community. The petascale HPC environment will enable investigations of computationally challenging problems that require computers operating at sustained speeds on actual research codes of 10^{15} floating point operations per second (petaflops) or that work with extremely large data sets on the order of 10^{15} bytes (petabytes).

Petascale HPC capabilities will permit researchers to perform simulations that are intrinsically multi-scale or that involve multiple simultaneous reactions, such as modeling the interplay between genes, microbes, and microbial communities and simulating the interactions between the ocean, atmosphere, cryosphere and biosphere in Earth systems models. In addition to addressing the most computationally challenging demands of science and engineering, new and improved HPC software services will make supercomputing platforms supported by NSF and other partner organizations more efficient, more accessible, and easier to use.

NSF will support the deployment of a well-engineered, scalable, HPC infrastructure designed to evolve as science and engineering research needs change. It will include a sufficient level of diversity, both in architecture and scale of deployed HPC systems, to realize the research and education goals of the broad science and engineering community. NSF's HPC investments will be complemented by its simultaneous investments in data analysis and visualization facilities essential to the effective transformation of data products into information and knowledge.

The following principles will guide the agency's FY 2006 through FY 2010 investments.

- Science and engineering research and education priorities will drive HPC investments.
- Collaborative activities involving science and engineering researchers and private sector organizations are needed to ensure that HPC systems and services are optimally configured to support petascale scientific computing.
- Researchers and educators require access to reliable, robust, production-quality HPC resources and services.
- HPC-related research and development advances generated in the public and private sectors, both domestic and foreign, must be leveraged to enrich HPC capabilities.
- The development, implementation and annual update of an effective multi-year HPC. strategy is crucial to the timely introduction of research and development outcomes and innovations in HPC systems, software and services.

NSF's implementation plan to create a petascale environment includes the following three interrelated components:

1). Specification, Acquisition, Deployment and Operation of Science-Driven HPC Systems Architectures

An effective computing environment designed to meet the computational needs of a range of science and engineering applications will include a variety of computing systems with complementary performance capabilities. By 2010, the petascale computing environment available to the academic science and engineering community is likely to consist of: (i) a significant number of systems with peak performance in the 1-50 teraflops range, deployed and supported at the local level by individual campuses and other research organizations; (ii) multiple systems with peak performance of 100+ teraflops that support the work of thousands of researchers nationally; and, (iii) at least one system in the 1-10 petaflops range that supports a more limited number of projects demanding the highest levels of computing performance. All NSF-deployed systems will be appropriately balanced and will include core computational hardware, local storage of sufficient capacity, and appropriate data analysis and visualization capabilities. Chapters 3 and 4 in this document describe the complementary investments necessary to provide effective data analysis and visualization capabilities, and to integrate HPC resources into a comprehensive national CI environment to improve both accessibility and usability.

Over the FY 2006-2010 period, NSF will focus on HPC system acquisitions in the 100 teraflops to 10 petaflops range, where strategic investments on a national scale are necessary to ensure international leadership in science and engineering. Since different science and engineering codes may achieve optimal performance on different HPC architectures, it is likely that by 2010 the NSF-supported HPC environment will include both loosely-coupled and tightly coupled systems, with several different memory models.

To address the challenge of providing the research community with access to a range of HPC architectures within a constrained budget, a key element of NSF's strategy is to participate in resource-sharing with other federal agencies. A strengthened interagency partnership will focus, to the extent practicable, on ensuring shared access to federal leadership-class resources with different architectures, and on the coordination of investments in HPC system acquisition and operation. The Department of Energy's Office of Science and National Nuclear Security Administration have very active programs in leadership computing. The Department of Defense's (DOD's) High Performance Computing Modernization Office (HPCMOD) provisions HPC resources and services for the DOD science and engineering community, while NASA is deploying significant computing systems also of interest to NSF PIs. To capitalize on these common interests, NSF will work toward the creation of a Leadership Computing Council as proposed by Simon et al.[1], to include representatives from all federal agencies with a stake in science and engineering-focused HPC. As conceived, the Leadership Computing Council will make coordinated and collaborative investments in science-driven hardware architectures, will increase the diversity of architectures of leadership class systems available to researchers and educators around the country, will promote sharing of lessons learned, and will provide a richer HPC environment for the user communities supported by each agency.

Strong partnerships involving universities, industry and government are also critical to success. In addition to leveraging the promise of Phase III of the DARPA-sponsored High Productivity

[1] Simon *et al.,* "Science-Driven System Architecture: A New Process for Leadership Class Computing," Journal of the Earth Simulator, pages 1-9, Vol. 2, January 2005.

Computing Systems (HPCS) program[2] in which NSF is a mission partner, the agency will establish a discussion and collaboration forum for scientists and engineers - including computational and computer scientists and engineers - and HPC system vendors, to ensure that HPC systems are optimally configured to support state-of-the-art scientific computing. On the one hand, these discussions will keep NSF and the academic community informed about new products, product roadmap and technology challenges at various vendor organizations. On the other, they will provide HPC system vendors with insights into the major concerns and needs of the academic science and engineering community. These activities will lead to better alignment between applications and hardware both by influencing algorithm design and by influencing system integration.

NSF will also promote resource sharing between and among academic institutions to optimize the accessibility and use of HPC assets deployed and supported at the campus level. This will be accomplished through development of a shared governance structure that includes relevant HPC stakeholders.

2). Development and Maintenance of Supporting Software: New Design Tools, Performance Modeling Tools, Systems Software, and Fundamental Algorithms.

Many of the HPC software and service building blocks in scientific computing are common to a number of science and engineering applications. A supporting software and service infrastructure will accelerate the development of the scientific application codes needed to solve challenging scientific problems, and will help insulate these codes from the evolution of future generations of HPC hardware.

Supporting software services include the provision of intelligent development and problem-solving environments and tools. These are designed to provide improvements in ease of use, reusability of modules, and portable performance. Tools and services that take advantage of commonly-supported software tools can deliver similar work environments across different HPC platforms, greatly reducing the time-to-solution of computationally-intensive research problems by permitting local development of research codes that can then be rapidly transferred to, or incorporate services provided by, larger production environments. These tools and workflows built from collections of such tools can also be packaged for more general use. Applications scientists and engineers will also benefit from the development of new tools and approaches to debugging, performance analysis, and performance optimization.

Specific applications depend on a broad class of numerical and non-numerical algorithms that are widely used by many applications; for example, linear algebra, fast spectral transforms, optimization algorithms, multi-grid methods, adaptive mesh refinement, symplectic integrators, and sorting and indexing routines. To date, improved or new algorithms have been important contributors to performance improvements in science and engineering applications, the development of multi-grid solvers for elliptic partial differential equations being a prime example. Innovations in algorithms will have a significant impact on the performance of applications software. The development of algorithms for different architectural environments is an essential component of the effort to develop portable, scalable, applications software. Other important software services include libraries for communications services, such as MPI and OpenMP.

[2] The DARPA High Productivity Computing Systems is focused on providing a new generation of economically viable high productivity computing systems. HPCS program researchers have initiated a fundamental reassessment of how performance, programmability, portability, robustness and ultimately, productivity in the HPC domain are defined and measured.

The development and deployment of operating systems and compilers that scale to hundreds of thousands of processors are also necessary. They must provide effective fault-tolerance and must effectively insulate users from parallelization, latency management and thread management issues. To test new developments at large scales, operating systems and kernel researchers and developers must have access to the infrastructure necessary to test their developments at scale.

NSF will support Software Services Providers (SSPs) to develop this supporting software infrastructure. SSPs will be *individually and collectively* responsible for: applied research and development of supporting technologies; harvesting promising supporting software technologies from the research communities; performing scalability/reliability tests to explore software viability; developing, hardening and maintaining software where necessary; and facilitating the transition of commercially viable software into the private sector. SSPs will also support general software engineering consulting services for science and engineering applications, and will provide software engineering consulting support to individual researchers and to research and education teams as necessary.

SSPs will be responsible for ensuring software interoperability with other components of the cyberinfrastructure software stack, such as those generated to provide Data, Data Analysis and Visualization services, and Cyber-services and Virtual Organization capabilities – see Chapters 3 and 4 in this document. This will be accomplished through the creation and utilization of appropriate software test harnesses and will ensure that sufficient configuration controls are in place to support the range of HPC platforms used by the research and education community. The applications community will identify needed improvements in supporting software and will provide input and feedback on the quality of services provided by SSPs.

To guide the evolution of the SSP program, NSF will establish an HPC Software Services Council that includes representatives from academe, federal agencies and private sector organizations, including 3rd party and system vendors. The HPC Software Services Council will provide input on the strengths, weaknesses, opportunities and gaps in the software services currently available to the science and engineering research and education communities.

To minimize duplication of effort and to optimize the value of HPC services provided to the science and engineering community, NSF's investments will be coordinated with those of other agencies. DOE currently invests in software infrastructure centers through the Scientific Discovery through Advanced Computing (SciDAC) program, while DARPA's investments in the HPCS program contribute significant systems software and hardware innovations. NSF will seek to leverage and add value to ongoing DOE and DARPA efforts in this area.

3). Development and Maintenance of Portable, Scalable Applications Software

Today's microprocessor-based terascale computers place considerable demands on our ability to manage parallelism, and to deliver large fractions of peak performance. As the agency seeks to create a petascale computing environment, it will embrace the challenge of developing or converting key application codes to run effectively on new and evolving system architectures.

Over the FY 2006 through 2010 period, NSF will make significant new investments in the development, hardening, enhancement and maintenance of scalable applications software,

including community models, to exploit the full potential of current terascale and future petascale systems architectures. The creation of well-engineered, easy-to-use software will reduce the complexity and time-to-solution of today's challenging scientific applications. NSF will promote the incorporation of sound software engineering approaches in existing widely-used research codes and in the development of new research codes. Multidisciplinary teams of researchers will work together to create, modify and optimize applications for current and future systems using performance modeling tools and simulators.

Since the nature and genesis of science and engineering codes varies across the research landscape, a successful programmatic effort in this area will weave together several strands. A new activity will be designed to take applications that have the potential to be widely used within a community or communities, to harden these applications based on modern software engineering practices, to develop versions for the range of architectures that scientists wish to use them on, to optimize them for modern HPC architectures, and to provide user support.

++

CHAPTER 3
PLAN FOR DATA, DATA ANALYSIS, AND VISUALIZATION (2006-2010)

I. A WEALTH OF SCIENTIFIC OPPORTUNITIES AFFORDED BY DIGITAL DATA

Science and engineering research and education have become increasingly data-intensive, as a result of the proliferation of digital technologies and pervasive networks through which data are collected, generated, shared and analyzed. Worldwide, scientists and engineers are producing, accessing, analyzing, integrating and storing terabytes of digital data daily through experimentation, observation and simulation. Moreover, the dynamic integration of data generated through observation and simulation is enabling the development of new scientific methods that adapt intelligently to evolving conditions to reveal new understanding. The enormous growth in the availability and utility of scientific data is increasing scholarly research productivity, accelerating the transformation of research outcomes into products and services, and enhancing the effectiveness of learning across the spectrum of human endeavor.

New scientific opportunities are emerging from increasingly effective data organization, access and usage. Together with the growing availability and capability of tools to mine, analyze and visualize data, the emerging data cyberinfrastructure is revealing new knowledge and fundamental insights. For example, analyses of DNA sequence data are providing remarkable insights into the origins of man, are revolutionizing our understanding of the major kingdoms of life, and are revealing stunning and previously unknown complexity in microbial communities. Sky surveys are changing our understanding of the earliest conditions of the universe and providing comprehensive views of phenomena ranging from black holes to supernovae. Researchers are monitoring socio-economic dynamics over space and time to advance our understanding of individual and group behavior and their relationship to social, economic and political structures. Using combinatorial methods, scientists and engineers are generating libraries of new materials and compounds for health and engineering, and environmental scientists and engineers are acquiring and analyzing streaming data from massive sensor networks to understand the dynamics of complex ecosystems.

In this dynamic research and education environment, science and engineering data are constantly being collected, created, deposited, accessed, analyzed and expanded in the pursuit of new knowledge. In the future, U.S. international leadership in science and engineering will increasingly depend upon our ability to leverage this reservoir of scientific data captured in digital form, and to transform these data into information and knowledge aided by sophisticated data mining, integration, analysis and visualization tools.

This chapter sets forth a framework in which NSF will work with its partners in science and engineering – public and private sector organizations both foreign and domestic representing data producers, scientists, engineers, managers and users alike – to address data acquisition, access, usage, stewardship and management challenges in a comprehensive way.

II. DEFINITIONS

A. Data, Metadata and Ontologies

In this document, "data" and "digital data" are used interchangeably to refer to data and information stored in digital form and accessed electronically.

- *Data*. For the purposes of this document, data are any and all complex data entities from observations, experiments, simulations, models, and higher order assemblies, along with the associated documentation needed to describe and interpret the data.
- *Metadata*. Metadata are a subset of data, and are data about data. Metadata summarize data content, context, structure, inter-relationships, and provenance (information on history and origins). They add relevance and purpose to data, and enable the identification of similar data in different data collections.
- *Ontology*. An ontology is the systematic description of a given phenomenon, often includes a controlled vocabulary and relationships, captures nuances in meaning and enables knowledge sharing and reuse.

B. Data Collections

This document adopts the definition of data collection types provided in the NSB report on Long-Lived Digital Data Collections[3], where data collections are characterized as being one of three functional types:

- *Research Collections.* Authors are individual investigators and investigator teams. Research collections are usually maintained to serve immediate group participants only for the life of a project, and are typically subjected to limited processing or curation. Data may not conform to any data standards.
- *Resource Collections.* Resource Collections are authored by a community of investigators, often within a domain of science or engineering, and are often developed with community-level standards. Budgets are often intermediate in size. Lifetime is between the mid- and long-term.
- *Reference Collections.* Reference collections are authored by and serve large segments of the science and engineering community, and conform to robust, well-established, comprehensive standards, which often lead to a universal standard. Budgets are large and often derived from diverse sources with a view to indefinite support.

Boundaries between the types are not rigid and collections originally established as research collections may evolve over time to become resource and/or reference collections. In this document, the term data collection is construed to include one or more databases and their relevant technological implementation. Data collections are managed by organizations and individuals with the necessary expertise to structure them and to support their effective use.

III. DEVELOPING A COHERENT DATA CYBERINFRASTRUCTURE IN A COMPLEX, GLOBAL CONTEXT

Since data and data collections are owned or managed by a wide range of communities, organizations and individuals around the world, NSF must work in an evolving environment constantly being shaped by developing international and national policies and treaties,

[3] Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century, NSB-05-40

community-specific policies and approaches, institutional-level programs and initiatives, individual practices, and continually advancing technological capabilities.

At the international level, a number of nations and international organizations have already recognized the broad societal, economic, and scientific benefits that result from open access to science and engineering digital data. In 2004 more than thirty nations, including the United States, declared their joint commitment to work toward the establishment of common access regimes for digital research data generated through public funding[4]. Since the international exchange of scientific data, information and knowledge promises to significantly increase the scope and scale of research and its corresponding impact, these nations are working together to define the implementation steps necessary to enable the global science and engineering system.

The U.S. community is engaged through the Committee on Data for Science and Technology (CODATA). The USNC/CODATA is working with international CODATA partners, including the International Council for Science (ICSU), the International Council for Scientific and Technical Information (ICTSI), the World Data Centers (WDCs) and others, to accelerat the development of a global open-access scientific data and information resource, through the construction of an online "open access knowledge environment", as well as through targeted projects. Among other things, the Global Information Commons for Science. Initiative will facilitate reuse of publicly-funded scientific data and information, as well as cooperative sharing of research materials and tools among researchers,and encourage and coordinate the efforts of many stakeholders in the world's diverse science and engineering community to achieve these objectives.

A number of international science and engineering communities have also been developing data management and curation approaches for reference and resource collections. For example, the international Consultative Committee for Space Data Standards (CCSDS) defined an archive reference model and service categories for the intermediate and long-term storage of digital data relevant to space missions. This effort produced the Open Archival Information System (OAIS), now adopted as the "de facto" standard for building digital archives, and evidence that a community-focused activity can have much broader impact than originally intended. In another example, the Inter-University Consortium for Political and Social Research (ICPSR) - a membership-based organization with over 500 member colleges and universities around the world - maintains and provides access to a vast archive of social science data. ICPSR serves as a content management organization, preserving relevant social science data and migrating them to new storage media as technology changes, and also provides user support services. ICPSR recently announced plans to establish an international standard for social science documentation. Similar activities in other communities are also underway. Clearly, NSF must maintain a presence in, support, and add value to these ongoing international discussions and activities.

Activities on an international scale are complemented by activities within nation states. In the United States, a number of organizations and communities of practice are exploring mechanisms to establish common approaches to digital data access, management and curation. For example, the Research Library Group (RLG – a not for profit membership organization representing libraries, archives and museums) and the U.S. National Archives and Records Administration (NARA – a sister agency whose mission is to provide direction and

[4] http://www.oecd.org/document/0,2340,en_2649_34487_25998799_1_1_1_1,00.html

assistance to Federal agencies on records management) are producing certification requirements for establishing and selecting reliable digital information repositories. RLG and NARA intend their results to be standardized via the International Organization of Standardization (ISO) Archiving Series, and may impact all data collections types. The NIH National Center for Biotechnology Information plays an important role in the management of genome data at the national level, supporting public databases, developing software tools for analyzing data, and disseminating biomedical information.

At the institutional level, colleges and universities are developing approaches to digital data archiving, curation, and analysis. They are sharing best practices to develop digital libraries that collect, preserve, index and share research and education material produced by faculty and other individuals within their organizations. The technological implementations of these systems are often open-source and support interoperability among their adopters. University-based research libraries and research librarians are positioned to make significant contributions in this area, where standard mechanisms for access and maintenance of scientific digital data may be derived from existing library standards developed for print material. These efforts are particularly important to NSF as the agency considers the implications of not just making all data generated with NSF funding broadly accessible, but of also promoting the responsible organization and management of these data such that they are widely usable.

IV. PLAN OF ACTION

Motivated by a vision in which science and engineering digital data are routinely deposited in well-documented form, are regularly and easily consulted and analyzed by specialists and non-specialists alike, are openly accessible while suitably protected, and are reliably preserved, NSF's five-year goal is twofold:

- To catalyze the development of a system of science and engineering data collections that is open, extensible and evolvable; and
- To support development of a new generation of tools and services facilitating data mining, integration, analysis, and visualization essential to turning data into new knowledge and understanding.

The resulting *national digital data framework* will be an integral component in the national cyberinfrastructure framework described in this document, and will consist of a range of data collections and managing organizations, networked together in a flexible technical architecture using standard, open protocols and interfaces, and designed to contribute to the emerging global information commons. It will be simultaneously local, regional, national and global in nature, and will evolve as science and engineering research and education needs change and as new science and engineering opportunities arise. Widely accessible tools and services will permit scientists and engineers to access and manipulate these data to advance the science and engineering frontier.

In print form, the preservation process is handled through a system of libraries and other repositories throughout the country and around the globe. Two features of this print-based system make it robust. First, the diversity of business models deriving support from a variety of sources means that no single entity bears sole responsibility for preservation, and the system is resilient to changes in any particular sector. Second, there is overlap in the collections, and redundancy of content reduces the potential for catastrophic loss of information.

The *national data framework* is envisioned to provide an equally robust and diverse system for digital data management and access. It will: promote interoperability between data collections supported and managed by a range of organizations and organization types; provide for appropriate protection and reliable long-term preservation of digital data; deliver computational performance, data reliability and movement through shared tools, technologies and services; and accommodate individual community preferences. The agency will also develop a suite of coherent data policies that emphasize open access and effective organization and management of digital data, while respecting the data needs and requirements within science and engineering domains.

The following principles will guide the agency's FY 2006 through FY 2010 investments.

- Science and engineering research and education opportunities and priorities will motivate NSF investments in data cyberinfrastructure.
- Science and engineering data generated with NSF funding will be readily accessible and easily usable, and will be appropriately, responsibly and reliably preserved.
- Broad community engagement is essential in the prioritization and evaluation of the utility of scientific data collections, including the possible evolution between research, resource and reference collection types.
- Continual exploitation of data in the creation of new knowledge requires that investigators have access to the tools and services necessary to locate and access relevant data, and understand its structure sufficiently to be able to interpret and (re)analyze what they find.
- The establishment of strong, reciprocal, international, interagency and public-private partnerships is essential to ensure all stakeholders are engaged in the stewardship of valuable data assets. Transition plans, addressing issues such as media, stewardship, and standards, will be developed for valuable data assets, to protect data and to assure minimal disruption to the community during transition periods.
- Mechanisms will be created to share data stewardship best practices between nations, communities, organizations and individuals.
- In light of legal, ethical and national security concerns associated with certain types of data, mechanisms essential to the development of both statistical and technical ways to protect privacy and confidentiality will be supported.

A. A Coherent Organizational Framework - Data Collections and Managing Organizations

To date, challenges associated with effective stewardship and preservation of scientific data have been more tractable when addressed through communities of practice that may derive support from a range of sources. For example, NSF supports the Incorporated Research Institutions for Seismology (IRIS) consortium to manage seismology data. Jointly with NIH and DOE, the agency supports the Protein Data Bank to manage data on the three-dimensional structures of proteins and nucleic acids. Multiple agencies support the University Consortium for Atmospheric Research, an organization that has provided access to atmospheric and oceanographic data sets, simulations, and outcomes extending back to the 1930s through the National Center for Atmospheric Research.

Existing collections and managing organization models reflect differences in culture and practice within the science and engineering community. As community proxies, data collections and their managing organizations can provide a focus for the development and dissemination of

appropriate standards for data and metadata content and format, guided by an appropriate community-defined governance approach. This is not a static process, as new disciplinary fields and approaches, data types, organizational models and information strategies inexorably emerge. This is discussed in detail in the Long-Lived Digital Data Collections report of the National Science Board.

Since data are held by many Federal agencies, commercial and non-profit organizations, and international entities, NSF will foster the establishment of interagency, public-private and international consortia charged with providing stewardship for digital data collections to promote interoperability across data collections. The agency will work with the broad community of science and engineering producers, managers, scientists and users to develop a common conceptual framework. A full range of mechanisms will be used to identify and build upon common ground across domain communities and managing organizations, engaging all stakeholders. Activities will include: the support of new projects; development and implementation of evaluation and assessment criteria that, amongst other things, reveal lessons learned across communities; support of community and inter-community workshops; and the development of strong partnerships with other stakeholder organizations. Stakeholders in these activities include data authors, data managers, data scientists and engineers, and data users representing a diverse range of communities and organizations, including universities and research libraries, government agencies, content management organizations and data centers, and industry.

To identify and promote lessons learned across managing organizations, NSF will continue to promote the coalescence of appropriate collections with overlapping interests, approaches, and services. This reduces data-driven fragmentation of science and engineering domains. Progress is already being made in some areas. For example, NSF has been working with the environmental science and engineering community to promote collaboration across disciplines ranging from ecology and hydrology to environmental engineering. This has resulted in the emergence of common cyberinfrastructure elements and new interdisciplinary science and engineering opportunities.

B. Developing A Flexible Technological Architecture

From a technological perspective, the *national data framework* must provide for reliable preservation, access, analysis, interoperability, and data movement, possibly using a web or grid services distributed environment. The architecture must use standard open protocols and interfaces to enable the broadest use by multiple communities. It must facilitate user access, analysis and visualization of data, addressing issues such as authentication, authorization and other security concerns, and data acquisition, mining, integration, analysis and visualization. It must also support complex workflows enabling data discovery. Such an architecture can be visualized as a number of layers providing different capabilities to the user, including data management, analysis, collaboration tools, and community portals. The connections among these layers must be transparent to the end user, and services must be available as modular units responsive to individual or community needs. The system is likely to be implemented as a series of distributed applications and operations supported by a number of organizations and institutions distributed throughout the country. It must provide for the replication of data resources to reduce the potential for catastrophic loss of digital information through repeated cycles of systems migration and all other causes since, unlike printed records, the media on which digital data are stored and the structures of the data are relatively fragile.

High quality metadata, which summarize data content, context, structure, inter-relationships, and provenance (information on history and origins), are critical to successful information management, annotation, integration and analysis. Metadata take on an increasingly important role when addressing issues associated with the combination of data from experiments, observations and simulations. In these cases, product data sets require metadata that describe, for example, relevant collection techniques, simulation codes or pointers to archived copies of simulation codes, and codes used to process, aggregate or transform data. These metadata are essential to create new knowledge and to meet the reproducibility imperative of modern science. Metadata are often associated with data via markup languages, representing a consensus around a controlled vocabulary to describe phenomena of interest to the community, and allowing detailed annotations of data to be embedded within a data set. Because there is often little awareness of markup language development activities within science and engineering communities, energy is expended reinventing what could be adopted or adapted from elsewhere. Scientists and engineers therefore need access to tools and services that help ensure that metadata are automatically captured or created in real-time.

Effective data analysis tools apply computational techniques to extract new knowledge through a better understanding of the data, its redundancies and relationships, by filtering extraneous information and by revealing previously unseen patterns. For example, the Large Hadron Collider at CERN generates such massive data sets that the detection of both expected events, such as the Higgs boson, and unexpected phenomena requires the development of new algorithms, both to manage data and to analyze it. Algorithms and their implementations must be developed for statistical sampling, for visualization, to enable the storage, movement and preservation of enormous quantities of data, and to address other unforeseen problems certain to arise.

Scientific visualization, including not just static images but also animation and interaction, leads to better analysis and enhanced understanding. Currently, many visualization systems are domain or application-specific and require a certain commitment to understand or learn to use. Making visualization services more transparent to the user lowers the threshold of usability and accessibility, and makes it possible for a wider range of users to explore or use a data collection. Analysis of data streams also introduces problems in data visualization and may require new approaches for representing massive, heterogeneous data streams.

Deriving knowledge from large data sets presents specific scaling problems due to the sheer number of items, dimensions, sources, users, and disparate user communities. The human ability to process visual information can augment analysis, especially when analytic results are presented in iterative and interactive ways. Visual analytics, the science of analytical reasoning enabled by interactive visual interfaces, can be used to synthesize the information content and derive insight from massive, dynamic, ambiguous, and even conflicting data. Suitable fully interactive visualizations help us absorb vast amounts of data directly, to enhance our ability to interpret and analyze otherwise overwhelming data. Researchers can thus detect the expected and discover the unexpected, uncovering hidden associations and deriving knowledge from information. As an added benefit, their insights are more easily and effectively communicated to others.

Creating and deploying visualization services requires new frameworks for distributed applications. In common with other cyberinfrastructure components, visualization requires easy-to-use, modular, extensible applications that capitalize on the reuse of existing technology. Today's successful analysis and visualization applications use a pipeline, component-based

system on a single machine or across a small number of machines. Extending to the broader distributed, heterogeneous cyberinfrastructure system will require new interfaces and work in fundamental graphics and visualization algorithms that can be run across remote and distributed settings.

To address this range of needs for data tools and services, NSF will work with the broad community to identify and prioritize needs. In making investments, NSF will complement private sector efforts, for example, those producing sophisticated indexing and search tools and packaging them as data services. NSF will support projects to: conduct applied research and development of promising, interoperable data tools and services; perform scalability/reliability tests to explore tool viability; develop, harden and maintain software tools and services where necessary; and, harvest promising research outcomes to facilitate the transition of commercially-viable software into the private sector. Data tools created and distributed through these projects will include not only access and ease-of-use tools, but tools to assist with data input, tools that maintain or enforce formatting standards, and tools that make it easy to include or create metadata in real time. Clearinghouses and registries from which all metadata, ontology, and markup language standards are provided, publicized, and disseminated must be developed and supported, together with the tools for their implementation. Data accessibility and usability will also be improved with the support of means for automating cross-ontology translation. Collectively, these projects will be responsible for ensuring software interoperability with other components of the cyberinfrastructure, such as those generated to provide High Performance Computing capabilities and to enable the creation of Cyber-services and Virtual Organizations.

The user community will work with tool providers as active collaborators to determine requirements and to serve as early users. Scientists, educators, students and other end users think of ways to use data and tools that the developers didn't consider, finding problems and testing recovery paths by triggering unanticipated behavior. Most importantly, an engaged set of users and testers will also demonstrate the scientific value of data collections. The value of repositories and their standards-based input and output tools arises from the way in which they enable discoveries. Testing and feedback are necessary to meet the challenges presented by current datasets that will only increase in size, often by orders of magnitude, in the future.

Finally, in addition to promoting the *use* of standards, tool and service developers will also promote the *stability* of standards. Continual changes to structure, access methods, and user interfaces, mitigate against ease of use, and against interoperability. Instead of altering a standard for a current need, developers will adjust their implementation of that need to fit within the standard. This is especially important for resource-limited research and education communities.

C. Developing and Implementing Coherent Data Policies

In setting priorities and making funding decisions, NSF is in a powerful position to influence data policy and management at research institutions. NSF's policy position on data is straightforward: all science and engineering data generated with NSF funding must be made broadly accessible and usable, while being suitably protected and preserved. Through a suite of coherent policies designed to recognize different data needs and requirements within communities, NSF will promote open access to well-managed data recognizing that this is essential to continued U.S. leadership in science and engineering.

In addition to addressing the technological challenges inherent in the creation of a *national data framework*, NSF's data policies will be redesigned to overcome existing sociological and cultural barriers to data sharing and access. Two actions are critical. NSF will conduct an inventory of existing policies, to bring them into accord across programs and to ensure coherence. This will lead to the development of a suite of harmonized policy statements supporting data open access and usability. NSF's actions will promote a change in culture such that the collection and deposition of all appropriate digital data and associated metadata become a matter of routine for investigators in all fields. This change will be encouraged through an NSF-wide requirement for data management plans in all proposals. These plans will be considered in the merit review process, and will be actively monitored post-award.

Policy and management issues in data handling occur at every level, and there is an urgent need for rational agency, national and international strategies for sustainable access, organization and use. Discussions at the interagency level on issues associated with data policies and practices will be supported by a new interagency working group on digital data recently proposed by NSF under the auspices of the Committee on Science of the National Science and Technology Council. This group will consider not only data challenges and opportunities discussed throughout this chapter, but especially the issues of cross-agency funding and access, the provision and preservation of data to and for other agencies, and monitoring agreements as agency imperatives change with time. Formal policies must be developed to include data quality and security, ethical and legal requirements, and technical and semantic interoperability issues, throughout the complete process from collection and generation to analysis and dissemination.

As already noted, many large science and engineering projects are international in scope, where national laws and international agreements directly affect data access and sharing practices. Differences arise over privacy and confidentiality, from cultural attitudes to ownership and use, in attitudes to intellectual property protection and its limits and exceptions, and because of national security concerns. Means by which to find common ground within the international community must continue to be explored.

++

CHAPTER 4
PLAN FOR CYBER-SERVICES AND VIRTUAL ORGANIZATIONS (2006-2010)

I. NEW FRONTIERS IN SCIENCE AND ENGINEERING THROUGH CYBER-SERVICES AND VIRTUAL ORGANIZATIONS

With access to state-of-the-art cyberinfrastructure services, many researchers and indeed entire fields of science and engineering now share access to world-class resources spanning experimental facilities and field equipment, distributed instrumentation, sensor networks and arrays, mobile research platforms, HPC systems, data collections, sophisticated analysis and visualization facilities, and advanced simulation tools. The convergence of information, grid, and networking technologies with contemporary communications now enables science and engineering communities to pursue their research and learning goals often in real-time and without regard to geography. In fact, the creation of end-to-end cyberinfrastucture systems – comprehensive cyber-services – by groups of individuals with common interests is permitting the establishment of Virtual Organizations (VOs) that are revolutionizing the conduct of science and engineering research and education. A VO is created by a group of individuals whose members and resources may be dispersed geographically yet who function as a coherent unit through the use of end-to-end cyberinfrastructure systems. These systems provide shared access to centralized or distributed resources and services, often in real-time. Such virtual organizations go by several names: collaboratory, co-laboratory, grid community, science gateway, science portal, etc.

During the past decade, NSF funding has catalyzed the creation of VOs across a broad spectrum of science and engineering fields[5], creating powerful and broadly accessible pathways to accelerate the transformation of research outcomes into knowledge, products, services, and new learning opportunities. With access to enabling tools and services, self-organizing communities can create end-to-end systems to: facilitate scientific workflows; collaborate on experimental designs; share information and knowledge; remotely operate instrumentation; run numerical simulations using computing resources ranging from desktop computers to HPC systems; archive, e-publish, access, mine, analyze, and visualize data; develop new computational models; and, deliver unique learning and workforce development activities.

Through VOs, researchers are exploring science and engineering phenomena in unprecedented ways. Scientists are now defining the structure of the North American lithosphere with an

[5] Representative VOs are described in Appendix D.

extraordinary level of detail through EarthScope, which integrates observational, analytical, telecommunications, and instrumentation technologies to investigate the structure and evolution of the North American continent, and the physical processes controlling earthquakes and volcanic eruptions. The Integrated Primate Biomaterials and Information Resource assembles, characterizes, and distributes high-quality DNA samples of known provenance with accompanying demographic, geographic, and behavioral information to advance understanding of human origins, the biological basis of cognitive processes, evolutionary history and relationships, and social structure, and provides critical scientific information needed to facilitate conservation of biological diversity. The Time-sharing Experiments for the Social Sciences (TESS) allows researchers to run their own studies on random samples of the population that are interviewed via the Internet. By allowing social scientists to collect original data tailored to their own hypotheses, TESS increases the precision with which social science advances can be made. Through the Network for Earthquake Engineering Simulation (NEES), the coupling of high performance networks, advanced computational tools, and 15 experimental facilities enables engineering researchers to test larger scale and more comprehensive structural and geomaterial systems to create new design methodologies and technologies for reducing losses during earthquakes and tsunamis.

This chapter calls for the establishment of a national cyberinfrastructure framework into which the HPC environment described in Chapter 2 and the national data framework described in Chpater 3 are integrated, enabling the development, deployment, evolution, and sustainable support of end-to-end cyberinfrastructure systems that will serve as transformative agents for 21st century science and engineering discovery and learning, promoting shared use and interoperability across all fields.

II. ESTABLISHING A FLEXIBLE, OPEN CYBERINFRASTRUCTURE FRAMEWORK

NSF's five-year goals are as follows:

- To catalyze the development, implementation and evolution of a functionally-complete national cyberinfrastructure that integrates both physical and cyberinfrastructure assets and services.
- To promote and support the establishment of world-class VOs that are secure, efficient, reliable, accessible, usable, pervasive, persistent and interoperable, and that are able to exploit the full range of research and education tools available at any given time.
- To support the development of common cyberinfrastructure resources, services, and tools that enable the effective, efficient creation and operation of end-to-end cyberinfrastructure systems for and across all science and engineering fields, nationally and internationally.

The following principles will guide NSF investments:

- NSF's investments in end-to-end cyberinfrastructure systems are driven by science and engineering opportunities and challenges.
- Common needs and opportunities are identified to improve the cost-effectiveness of NSF investments and to enhance interoperability.
- NSF investments promote equitable provision of and access to all types of physical, digital, and human resources to ensure the broadest participation of individuals with interest in science and engineering inquiry and learning.
- Existing projects and programs inform future investments, serving as a resource and knowledge base.

- End-to-end cyberinfrastructure systems are appropriately reliable, robust, and persistent such that end users can depend on them to achieve their research and education goals.
- NSF partners with relevant stakeholders, including academe, industry, other federal agencies, and other public and private sector organizations, both foreign and domestic.
- Tools and services are networked together in a flexible architecture using standard, open protocols and interfaces, designed to support the creation and operation of robust cyber-services and VOs across the scientific and engineering disciplines supported by NSF.

The cyberinfrastructure framework developed will integrate widely accessible, common cyberinfrastructure resources, services, and tools, such as those described in other chapters in this document, enabling individuals, groups and communities to efficiently design, develop, deploy, and operate flexible, customizable cyber-services and VOs to advance science and engineering.

In facilitating the creation and support of enabling cynerenvironments and virtual organizations, NSF will focus on three essential elements: the creation of a common technological framework that promotes seamless, secure integration across a wide range of shared, geographically-distributed resources; the establishment of an operational framework built on productive and accountable partnerships developed among system architects, developers, providers, operators, and end users who span multiple communities; and, the establishment of an effective assessment and evaluation plan that will inform the agency's ongoing investments in cyberinfrastructure for the foreseeable future.

A. Open Technological Framework

To facilitate the development of an open technological framework, NSF will support cyberinfrastructure **software service providers (SSPs)** to develop, integrate, deploy, and support reliable, robust, and interoperable software. Software essential to the creation of cyber-services and VOs encompasses a broad range of functionalities and services, including enabling middleware; domain-specific software and application codes; teleobservation and teleoperation tools to enable remote access to experimental facilities, instruments, and sensors; collaborative tools for experimental planning, execution, and post-analysis; workflow tools and processes; system monitoring and management; user support; web portals to open source simulation software and domain-specific community code repositories; and flexible user interfaces to enable discovery and learning.

Many of the projects listed in Appendix D have produced fundamental software and/or new integrated environments to support interdisciplinary research and education. NSF's strategy leverages this body of work, harvesting promising tools and technologies that have been developed to a research prototype stage, and further hardening, generalizing, and making them available for use by multiple individuals and/or communities. For example, many communities require access to services that build scientific workflows and rich orchestration tools, and to integrate the intensive computing and data capabilities described in Chapters 2 and 3, respectively, into collaborative and productive working environments. While work has already been done to develop and deploy components and packages of needed software through NSF and other support, existing software needs to be hardened, maintained, and evolved. New software must be developed as new uses and new user requirements continue to emerge.

NSF will also support development of cybersecurity tools and technologies. Cybersecurity pervades all aspects of end-to-end cyberinfrastructure systems and includes human, data, software, and facilities aspects. Security requires coordination, the development of trust, and rule setting through community governance. NSF will require awardees developing, deploying and supporting cyber-services and virtual organizations to develop and consistently apply robust cybersecurity policies and procedures, thereby promoting a conscientious approach to cybersecurity. The agency will support development of strong authentication and authorization technologies and procedures for individuals, groups, VOs, and other role-based identities that scale. In so doing, the agency will leverage research prototypes and new technologies produced in the Cyber Trust program, the U.S. government's flagship program in cybersecurity research and development.

Interoperable, open technology standards will be used as the primary mechanism to support the further development of interoperable, open, extensible cyber-services and VOs. Standards for data representation and communication, connection of computing resources, authentication, authorization, and access must be non-proprietary, internationally employed, accepted by multiple research and education communities. The use of standards creates economies of scale and scope for developing and deploying common resources, tools, software, and services that enhance the use of cyberinfrastructure in multiple science and engineering communities. This approach allows maximum interoperability and sharing of best practices. A standards-based approach will ensure that access cyberinfrastructure will be: independent of operating systems; ubiquitously accessible; and, open to large and small institutions. Together, web services and standard-oriented architectures are emerging as a standard framework for interoperability among various software applications on different platforms. They provide important characteristics such as standardized and well-defined interfaces and protocols, ease of development, and reuse of services and components, making them central facets of the cyberinfrastructure software ecosystem.

B. Operational Framework

NSF will also promote the development of partnerships to facilitate the sharing and integration of distributed technological components deployed and supported at national, international, regional, local, community, and campus levels. Significant resources already exist at the academic institution level. It is important to integrate such resources into the national cyberinfrastructure fabric. NSF will promote the integration of campus-based cyberinfrastructure through interactions with campus CIOs and their organizations as well as with departments and individual faculty, to achieve holistic end-to-end cyberinfrastructure systems.

In addition, NSF will establish partnerships with industry to develop, maintain, and share robust, production-quality software tools and services and to leverage commercially available software. NSF will engage commercial software providers to identify value propositions, to identify software needs specific to the research and education community, and to facilitate technology transfer to industry, where appropriate. Efforts to ensure that NSF cyberinfrastructure investments complement those of other federal agencies will be intensified.

With the increasing globalization of science and engineering, NSF will support international efforts of greatest strategic interest, and will facilitate U.S. researchers' collaboration with other international centers or peer teams. NSF will identify exemplars of international collaboration and partnerships that offer efficient and beneficial relationships and will build on these.

The cost-effective penetration of cyberinfrastructure into all aspects of research and education will require the full engagement of the broad science and engineering community. Incorporating the contributions from multiple communities and reconciling their interests is one of the major challenges ahead. Community proxies must be identified and empowered to find common interests to avoid duplication of effort and to minimize the balkanization of science and engineering.

C. Evaluation and Assessment

Cyberinfrastructure is dramatically altering the conduct of science and engineering research and education. Accordingly, studying the evolution and impact of cyberinfrastructure on the culture and conduct of research and education within and across communities of practice is essential. NSF will also support projects that study how ongoing and future cyberinfrastructure efforts might be informed by lessons learned and by the identification of promising practices. Amongst other things, NSF seeks to build a stronger foundation in our understanding of: how individuals, teams and communities most effectively interact with cyberinfrastructure; how to design the critical governance and management structures for the new types of organizations arising; and, how to improve the allocation of cyberinfrastucture resources and to design incentives for its optimal use. These types of activities will be essential to the agency's overall success. NSF will support studies of the evolution and impact of cyberinfrastructure on the culture and conduct of research and education within and across different research and education communities.

The rapidly evolving nature of cyberinfrastructure requires ongoing assessment of current and future user requirements. Comprehensive user assessments will be conducted and will include identification and evaluation of how the physical infrastructure, networking needs and capabilities, collaborative tools, software requirements, and data resources affect the ability of scientists and engineers to conduct transformative research and provide rich learning and workforce development environments. Other issues to be addressed include the degree to which cyberinfrastructure facilitates federated inquiry, interoperability, and the development of common standards.

++

CHAPTER 5
PLAN FOR LEARNING AND WORKFORCE DEVELOPMENT (2006-2010)

I. INTRODUCTION

Cyberinfrastructure moves us beyond the old-school model of teachers/students and classrooms/labs. Ubiquitous learning environments now will encompass classrooms, laboratories, libraries, galleries, museums, zoos, workplaces, homes and many other locations. Under this transformation, well-established components of education -- pre-school, K-12 and post-secondary -- become highly leveraged elements of a world where people learn as a routine part of life, throughout their lives.

Cyberinfrastructure is enabling powerful opportunities: i) to collaborate, ii) to model and visualize complex scientific and engineering concepts, iii) to create and discover scientific and educational resources for use in a variety of settings, both formal and informal, iv) to assess learning gains, and v) to customize or personalize learning environments. These changes demand a new level of technical competence in the science and engineering workforce and in our citizenry at large.

> *Learning through Interactive Visualizations and Simulations.* Imagine an interdisciplinary course in the design and construction of large public works projects, attracting student-faculty teams from different engineering disciplines, urban planning, environmental science, and economics; and from around the globe. To develop their understanding, the students combine relatively small self-contained digital simulations that capture both simple behavior and geometry to model more complex scientific and engineering phenomena. Modules share inputs and outputs and otherwise interoperate. These "building blocks" maintain sensitivity across multiple scales of phenomena. For example, component models of transportation subsystems from one site combine with structural and geotechnical models from other collections to simulate dynamic loading within a complex bridge and tunnel environment. Computational models from faculty research efforts are used to generate numerical data sets for comparison with data from physical observations of real transportation systems obtained from various (international) locations via access to remote instrumentation. Furthermore, learners explore influences on air quality and tap into the expertise of practicing environmental scientists through either real-time or asynchronous communication. This networked learning environment increases the impact and accessibility of all resources by allowing students to search for and discover content, to assemble curricular and learning modules from component pieces in a flexible manner, and to communicate and collaborate with others, leading to a deep change in the relationship between students and knowledge. Indeed, students experience the profound changes in the practice of science and engineering and the nature of inquiry that cyberinfrastructure provokes.

To realize these radical changes in the processes of learning and discovery, cyber-services also demand a new level of technical competence from the Nation's workforce and citizenry. Indeed, NSF envisions a spectrum of new learning needs and activities demanded by individuals, from future researchers, to members of the technical cyberinfrastructure workforce, to the citizen at large:

Future generations of research scientists and engineers. As cyberinfrastructure tools grow more accessible, students at the secondary school and undergraduate levels increasingly use them in their learning endeavors, in many cases serving as early adopters of emergent cyberinfrastructure. Already, these tools facilitate communication across disciplinary, organizational, and international and cultural barriers, and their use is characteristic of the new globally-engaged researcher. Moreover, the new tools and functionality of cyberinfrastructure are transforming the very nature of scientific inquiry and scholarship. New methods to observe and to acquire data, to manipulate it, and to represent it challenge the traditional discipline-based graduate curricula. Increasingly the tools of cyberinfrastructure must be incorporated within the context of disciplinary research. Indeed, these tools and approaches are helping to make possible new methods of inquiry that allow understanding in one area of science to promote insight in another, thus defining new interdisciplinary areas of research reflecting the complex nature of modern science and engineering problems. Furthermore, as data are increasingly "born digital," the ephemeral nature of data sources themselves raise new dimensions to the issues of preservation and stewardship.

Teachers and faculty. To employ the tools and capabilities of cyberinfrastructure-enabled learning environments effectively, teachers and faculty must also have continued professional development opportunities. For example, teachers and faculty must learn to use new assessment techniques and practices enabled by cyberinfrastructure, including the tailoring of feedback to the individual, and the creation of personalized portfolios of student learning that capture a record of conceptual learning gains over time. Undergraduate curricula must also be reinvented to exploit emerging cyberinfrastructure capabilities. The full engagement of students is vitally important since they are in a special position to inspire future students with the excitement and understanding of cyberinfrastructure-enabled scientific inquiry and learning.

Cyberinfrastructure career professionals. Ongoing attention must be paid to the education of the professionals who will support, deploy, develop, and design current and emerging cyberinfrastructure. For example, the increased emphasis on "data rich" scientific inquiry has revealed serious needs for "digital data management" or data curation professionals. Such careers may involve the development of new, hybrid degree programs that marry the study of library science with a scientific discipline. Similarly, the power that visualization and other presentation technologies bring to the interpretation of data may call for specialized career opportunities that pair the graphical arts with a science or engineering discipline.

Business and industry workforce. Cyberinfrastucture's impact on the conduct of business demands that members of the workforce have the capability at least to refresh if not also retool their skills. In some cases the maintenance of formal professional certifications to practice is a driver, and in other cases the need for continual workplace learning is driven by pressures to remain competitive and/or relevant to a sector's needs.

Adequate cyberinfrastructure must be present to support such intentional workforce development.

Citizens at large. Cyberinfrastructure extends the impact of science to citizens at large by enhancing communication about scientific inquiry and outcomes to the lay public. Such informal learning opportunities answer numerous needs including those of parents involved with their children's schooling and adults involved with community development needs that have scientific dimensions. Moreover, cyberinfrastructure enables lifelong learning opportunities as it supports the direct involvement by citizens in distributed scientific inquiry such as contributing to the digital sky survey.

Just as cyberinfrastructure changes the needs and roles of the individual learner, NSF also envisions it changing the organizational enterprise of learning. Two intertwined assumptions underlie this vision. Firstly, "online" will be the dominant operating mode for individuals, characterizing how individuals interact with educational resources and complementing how they interact with each other. Secondly, ubiquitous (or pervasive) computing will extend awareness of our physical and social environment, with embedded smart sensors and "device to device" communication becoming the norm. Moreover, the shift from wired to wireless will untether the learner from fixed formal educational settings and enable "on demand/on location" learning whether at home, in the field, in the laboratory, or at the worksite, locally or across the globe.

These conditions permit new learning organizations to form, raising in turn new research questions about the creation, operation, and persistence of communities of practice and learning. In such ***cyberlearning*** networks people will connect to learn with each other, even as they learn to connect with each other, to exploit increasingly shared knowledge and engage in participatory inquiry.

To support this vision of (massively) networked learning, cyberinfrastructure must be adaptive and agile; in short, a dynamic ecosystem that supports interactive, participatory modes of learning and inquiry, and that can respond flexibly to the infusion of new technology.

II. PRINCIPLES

To guide its decision-making, NSF has identified the following principles.

- Equitable and broad access to state-of-the-art cyber-services is essential.
- To achieve widespread use of cyberinfrastructure by science and engineering researchers, educators, and learners, efficient methods must exist to find, access and use cyberinfrastructure resources, tools, and services as well as the educational materials associated with them.
- The privacy, social, cultural, ethical and ownership issues associated with increasing use of cyberinfrastructure for learning, research and scholarship are addressed.
- Learning and workforce development opportunities contribute to cyberinfrastructure developments.
- Cyberinfrastructure developments will lead to new learning models necessary for lifelong learning in the distributed and networked learning environment.
- Leveraging cyberinfrastructure LWD activities and investments within NSF and by other agencies – national and international – are essential for enabling 21st century science and engineering.

- Scientists and engineers must be prepared to collaborate across disciplinary, institutional, geopolitical and cultural boundaries using cyberinfrastructure-mediated tools.

III. GOALS AND STRATEGIES

NSF's goals for Learning and Workforce Development are five-fold as indicated below:

1. Foster the broad deployment and utilization of cyberinfrastructurel-enabled learning and research environments.

Current and future generations of scientists, engineers and educators will utilize cyberinfrastructure-enabled learning and research environments for their formal and informal educational training, research projects, career development and life-long learning. Therefore, NSF will develop and implement strategies for their deployment and utilization. To do so, NSF will stimulate awareness of cyberinfrastructure-enabled learning and research environments for scientists, engineers and educators; enhance usability and adaptability of cyberinfrastructure-enabled learning and research environments for current and future generations of scientists, engineers and educators both nationally and internationally; promote new approaches to and integration of cyberinfrastructure-enabled learning and research environments for educational and research usage nationally and internationally; and foster incorporation of solutions for addressing privacy, ethical and social issues in cyberinfrastructure-enabled learning and research environments.

2. Support the development of new skills and professions needed for full realization of CI-enabled opportunities.

It is very likely that new disciplines will develop as a natural outgrowth of the advances in cyberinfrastructure. NSF will be an enabler in developing the workforce in these newly-formed disciplines. These disciplines could be as important as the relatively new disciplines of computer science, mathematical biology, genomics, environmental sciences and astrophysics are today. NSF will support mechanisms for development of new cyberinfrastructure-related curriculum at all levels; stimulate partnerships – domestic and international– between and among academic institutions and industrial cyberinfrastructure- professionals, and support the wide dissemination of "best practices" in cyberinfrastructure workforce development.

3. Promote broad participation of underserved groups, communities and institutions, both as creators and users of CI.

Cyberinfrastructure has the potential to enable a larger and more diverse set of individuals and institutions to participate in science and engineering education, research and innovation. To realize this potential, NSF will strategically design and implement programs that recognize the needs of those who might not have the means to utilize CI in science and engineering research and education. To do so, NSF will identify and address barriers to utilization of cyberinfrastructure tools and resources; promote the training of faculty particularly those in minority-serving institutions, predominantly undergraduate institutions and community colleges; and encourage programs to integrate innovative methods of teaching and learning using cybertools (particularly in inner-city, rural and remote classrooms), including taking advantage of international cyber-services to prepare a globally engaged workforce.

4. Stimulate new developments and continual improvements of cyberinfrastructure-enabled learning and research environments

In the dynamic environment of cyberinfrastructure-enabled learning and research, NSF will facilitate new developments as well as continuous improvements of the currently available tools and services, including those for education and training. NSF will support research that increases understanding of how students, teachers, scientists and engineers work and learn in a cyberinfrastructure rich environment, for example, interactive gaming, simulation, and modeling (as opposed to conventional instruction methods). The agency will: support the development of methods to embed relevant data collection and analysis tools in cyberinfrastructure-based environments in order to assess, e.g., satisfaction, usability, utility, productivity, etc., as well as the development of specific means for tracking student progress in content-based cyberinfrastructure learning; promote the development of technological solutions for addressing privacy, ethical and social issues in cyberinfrastructure-enabled learning and research environments; and stimulate partnerships both domestic and international to identify best practices in cyberinfrastructure enabled learning and research environments.

5. Facilitate cyberinfrastructure-enabled lifelong learning opportunities ranging from the enhancement of public understanding of science to meeting the needs of the workforce seeking continuing professional development.

Lifelong learning, through both formal and informal mechanisms, will be an essential part of the workforce of a cyberinfrastructure-enabled society. NSF can play a crucial role by promoting and sustaining programs that exploit existing resources and encourage creation of new resources in order to continually improve the science literacy of society in general and of the science and engineering workforce in particular. NSF will support mechanisms for professionals to continuously update their cyberinfrastructure skills and competencies; catalyze the development of new lifelong learning cyber-services; promote new knowledge communities and networks that take advantage of cyberinfrastructure to provide new learning experiences; support programs that bridge pre and post professional (lifelong) learning; and encourage programs that promote the public awareness and literacy in cyberinfrastructure.

++

APPENDIX A: REPRESENTATIVE REPORTS AND WORKSHOPS

Building a Cyberinfrastructure for the Biological Sciences; workshop held July 14-15, 2003; information available at http://research.calit2.net/cibio/archived/CIBIO_FINAL.pdf and http://research.calit2.net/cibio/report.htm

CHE Cyber Chemistry Workshop; workshop held October 3-5, 2004; information available at http://bioeng.berkeley.edu/faculty/cyber_workshop

Commission on Cyberinfrastructure for the Humanities and Social Sciences; sponsored by the American Council of Learned Societies; seven public information-gathering events held in 2004; report in preparation; information available at http://www.acls.org/cyberinfrastructure/cyber.htm

Community Climate System Model Strategic Business Plan (2003), 28pp; information available at http://www.ccsm.ucar.edu/management/busplan2004-2008.pdf

Community Climate System Model Science Plan 2004-2008 (2003)," 76pp; information available at http://www.ccsm.ucar.edu/management/sciplan2004-2008.pdf

Computation as a Tool for Discovery in Physics; report by the Steering Committee on Computational Physics: information available at http://www.nsf.gov/pubs/2002/nsf02176/start.htm

Cyberinfrastructure for the Atmospheric Sciences in the 21st Century; workshop held June 2004; information available at http://netstats.ucar.edu/cyrdas/report/cyrdas_report_final.pdf

Cyberinfrastructure for Engineering Research and Education; workshop held June 5 – 6, 2003; information available at http://www.nsf.gov/eng/general/Workshop/cyberinfrastructure/index.jsp

Cyberinfrastructure for Environmental Research and Education (2003*)*; workshop held October 30 – November 1, 2002; information available at http://www.ncar.ucar.edu/cyber/cyberreport.pdf

CyberInfrastructure (CI) for the Integrated Solid Earth Sciences (ISES) (June 2003); *workshop held on March 28-29, 2003;*, June 2003; information available at http://tectonics.geo.ku.edu/ises-ci/reports/ISES-CI_backup.pdf

Cyberinfrastructure and the Social Sciences (2005); workshop held March 15-17, 2005; information available at http://www.sdsc.edu/sbe/

Cyberinfrastructure needs for environmental observatories; information available at http://www.orionprogram.org/office/NSFCyberWkshp.html

Cyberlearning Workshop Series; workshops held Fall 2004 – Spring 2005 by the Computing Research Association (CRA) and the International Society of the Learning Sciences (ISLS); information available at http://www.cra.org/Activities/workshops/cyberlearning

Data Management for Marine Geology and Geophysics: Tools for Archiving, Analysis, and Visualization (2001); information available at http://humm.whoi.edu/DBMWorkshop/data_mgt_report.hi.pdf

Environmental Cyberinfrastucture Needs For Distributed Sensor Networks; workshop held August 12-14, 2003; information available at http://www.lternet.edu/sensor_report

Federal Plan for High-End Computing (2004); 72 pp; available at: http://www.ostp.gov/nstc/html/HECRTF-FINAL_051004.pdf

Geoinformatics: Building Cyberinfrastructure for the Earth Sciences (2004)*;* workshop held May 14 – 15, 2003; Kansas Geological Survey Report 2004-48; information available at http://www.geoinformatics.info

Geoscience Education and Cyberinfrastructure, Digital Library for Earth System Education, (2004); workshop held April 19-20, 2004; information available at http://www.dlese.org/documents/reports/GeoEd-CI.pdf

Getting Up to Speed: The Future of Supercomputing (2004). 308pp; available at: http://www.nap.edu/books/0309095026/html/ or http://www.sc.doe.gov/ascr/Supercomputing%20Prepub-Nov9v4.pdf)

High-Performance Computing Requirements for the Computational Solid Earth Sciences (2005)*;* 96 pp; available at: http://www.geo-prose.com/computational_SES.html.

Identifying Major Scientific Challenges in the Mathematical and Physical Sciences and their CyberInfrastructure Needs, workshop held April 21, 2004; information available at http://www.nsf.gov/attachments/100811/public/CyberscienceFinal4.pdf

Improving the effectiveness of U.S. Climate modeling, Commission on Geosciences, Environment and Resources (2001). National Academy Press, Washington, D.C., 144pp; information available at http://www.nap.edu/books/0309072573/html/

An Information Technology Infrastructure Plan to Advance Ocean Sciences (2002). 80 pp. available at http://www.geo-prose.com/oiti/index.html

Materials Research Cyberscience enabled by Cyberinfrastructure; workshop held June 17 – 19, 2004; information available at http://www.nsf.gov/mps/dmr/csci.pdf

Multi-disciplinary Workshop at the Interface of Cyber infrastructure, and Operations Research, with "Grand Challenges" in Enterprise-wide Applications in Design, Manufacturing and Services; workshop held August 31 - September 1, 2004; information available at https://engineering.purdue.edu/PRECISE/CI-OR/index.html

Multiscale Mathematics Initiative: A Roadmap; workshops held May 3-5, July 20-22, September 21-23, 2004; information available at www.sc.doe.gov/ascr/mics/amr/Multiscale%20Math%20Workshop%203%20-%20Report%20latest%20edition.pdf

NIH/NSF Spring 2005 Workshop on Visualization Research Challenges; workshop held on May 2-3, 2005; information available at http://www.sci.utah.edu/vrc2005/index.html

An Operations Cyberinfrastructure: Using Cyberinfrastructure and Operations Research to Improve Productivity in American Enterprises"; workshop held August 30 – 31, 2004; information available at http://www.optimization-online.org/OCI/OCI.doc; http://www.optimization-online.org/OCI/OCI.pdf

Planning for Cyberinfrastructure Software (2005); workshop held October 5 – 6, 2004; information available at www.nsf.gov/cise/sci/ci_workshop/index.jsp

Preparing for the Revolution: Information Technology and the Future of the Research University (2002); NRC Policy and Global Affairs, 80 pages; information available at http://www.nap.edu/catalog/10545.html

Polar Science and Advanced Networking: workshop held on April 24 - 26, 2003; sponsored by OPP/CISE; information available at http://www.polar.umcs.maine.edu

Recurring Surveys: Issues and Opportunities: workshop held March 28-29, 2003; information available at www.nsf.gov/sbe/ses/mms/nsf04_211a.pdf (2004)

Research Opportunities in CyberEngineering/CyberInfrastructure; workshop held April 22 - 23, 2004; information available at http://thor.cae.drexel.edu/~workshop/

Revolutionizing Science and Engineering Through Cyberinfrastructure: report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure; Daniel E. Atkins (Chair), January 2003; information available at http://www.nsf.gov/cise/sci/reports/atkins.pdf

Roadmap for the Revitalization of High-End Computing (200?); available at http://www.hpcc.gov/hecrtf-outreach/20040112_cra_hecrtf_report.pdf

Science-Based Case for Large-Scale Simulation; workshop held June 24-25, 2003; information available at http://www.pnl.gov/scales/docs/volume1_72dpi.pdf; http://www.pnl.gov/scales/docs/SCaLeS_v2_draft_toc.pdf

Supplement to the President's Budget for FY 2006; Report by the Subcommittee on Networking and Information Technology Research and Development (NITRD), February 2005; information available at http://www.nitrd.gov

Trends in IT Infrastructure in the Ocean Sciences (2004); workshop held May 21-23, 2003; information available at http://www.geo-prose.com/oceans_iti_trends/oceans_iti_trends_rpt.pdf

APPENDIX B: CHRONOLOGY OF NSF IT INVESTMENTS

NSF's early investments in what has now become known as cyberinfrastructure date back almost to the agency's inception. In the 1960's and 1970's, the agency supported a number of campus-based computing facilities. As computational methodologies became increasingly essential to the research endeavor, the science and engineering community began to call for NSF investments in specialized, higher capability computing facilities that would meet the computational needs of the broad national community. As a consequence, NSF's Supercomputer Centers program was initiated in 1985 through the agency's support of five academic-based supercomputer centers.

During the 1980's, academic-based networking activities also flourished. Networking technologies were expected to improve the effectiveness and efficiency of researchers and educators, providing enhanced, easier access to computer resources and more effective transfer and sharing of information and knowledge. After demonstrating the potential of CSNET in linking computer science departments, NSF moved on to develop the high-speed backbone, called NSFNET, with the five supercomputer centers supported under the Supercomputer Centers program and the National Center for Atmospheric Research becoming the first nodes on the backbone. NSF support also encouraged the development of regional networks to connect with the backbone NSFNET, thereby speeding the adoption of networking technologies on campuses around the country. In 1995, in partnership with MCI, NSF catalyzed support of the vBNS permitting advanced networking research and the development of novel scientific applications. A few years later, we established the NSF Middleware Initiative, focused on the development of advanced networking services to serve the evolving needs of the science and engineering community.

In the early to mid-1990's, informed by both the Branscomb and the Hayes Reports, NSF consolidated its support of national computing facilities in the establishment of the Partnerships for Advanced Computational Infrastructure (PACI) program. Two partnerships were established in 1997, together involving nearly 100 partner institutions across the country in efforts to make more efficient use of high-end computing in all areas of science and engineering. The partnerships have been instrumental in fostering the maturation of cyberinfrastructure and its widespread adoption by the academic research and education community, and by industry.

Also in the early 1990's, NSF as part of the U.S. High-Performance Computing and Communications (HPCC) program, began to support larger-scale research and education-focused projects pursuing what became known as "grand challenges." These HPCC projects joined scientists and engineers, computer scientists and state-of-the-art cyberinfrastructure technologies to tackle important problems in science and engineering whose solution could be advanced by applying cyberinfrastructure techniques and resources. First coined by the HPCC program, the term "grand challenge" has been widely adopted in many science and engineering fields to signify an overarching goal that requires a large-scale, concerted effort.

During the 1990's, the penetration of increasingly affordable computing and networking technologies on campuses was also leading to the creation of what would become mission-critical, domain-specific cyberinfrastructure. For example, in the mid 1990's the earthquake engineering community began to define what would become the Network for Earthquake Engineering Simulation, one of many significant cyberinfrastructure projects in NSF's portfolio today.

In 1999, the President's Information Technology Advisory Committee (PITAC) released the seminal report *ITR-Investing in our Future*, prompting new and complementary NSF investments in CI projects, such the Grid Physics Network (GriPhyN) and international Virtual Data Grid Laboratory (iVDGL) and the Geosciences Network, known as GEON. Informed by the PITAC report, NSF also created an MREFC project entitled Terascale Computing Systems that began its construction phase in FY 2000 and ultimately created the Extensible Terascale Facility – now popularly known as the Teragrid. Teragrid entered its production phase in October 2004 and represents one of the largest, fastest, most comprehensive distributed cyberinfrastructures for science and engineering research and education.

In 2001, NSF charged an Advisory Committee for Cyberinfrastructure under the leadership of Dr. Dan Atkins, to evaluate the effectiveness of PACI and to make recommendations for future NSF investments in cyberinfrastructure. The Atkins Committee, as it became popularly known, recommended support for the two Partnership lead sites through the end of their original PACI cooperative agreements. In October 2004, following merit review, the National Science Board (NSB) endorsed funding of those sites through the end of FY 2007.

Through 2005, in addition to the groups already cited, a number of prestigious groups have made recommendations that continue to inform the agency's cyberinfrastructure planning including the High-End Computing Revitalization Task Force, the PITAC Subcommittee on Computational Science, and the NRC Committee on the Future of Supercomputing.

APPENDIX C: MANAGEMENT OF CYBERINFRASTRUCTURE

NSF has nurtured the growth of what is now called cyberinfrastructure for a number of decades. In recent years, the Directorate for Computer and Information Science and Engineering (CISE) has been responsible for the provision of national supercomputing infrastructure for the academic community. In addition, the Directorate was instrumental in the creation of what ultimately became known as the Internet. During this incubation period, the management of CI was best provided by those also responsible for the research and development of related CI technologies.

Over the years, the penetration and impact of computing and networking on campuses has been extensive, and has led to the creation of many disciplinary-specific or community-specific CI projects and activities. Today, CI projects are supported by all NSF Directorates and Offices. Because of the growing scope of investment and variability in needs among users in the broad science and engineering community, it has become clear that effective CI development and deployment now requires the collective leadership of NSF senior management. This leadership will be provided by a Cyberinfrastructure Council chaired by the NSF Director and comprised of the NSF Deputy Director, the Assistant Directors of NSF's Directorates (BIO, CISE, GEO, EHR, ENG, MPS, and SBE) and the Heads of the Office of International Science and Engineering, Office of Polar Programs, and the recently established Office of Cyberinfrastructure (OCI). The Cyberinfrastructure Council has been meeting regularly since May 2005, and OCI was established in the Office of the Director on July 22, 2005.

CISE will continue to be responsible for a broad range of programs that address the Administration's priorities for fundamental research and education in computing, representing more than 85% of the overall federal investment in university-based basic research.

Appendix D: Representative Cyber-services and Virtual Organizations

Collaboratories:

- **CoSMIC: Combinatorial Sciences and Materials Informatics Collaboratory** http://mse.iastate.edu/cosmic-imi/overview.html An international research and education center promoting the use of informatics and combinatorial experimentation for materials discovery and design. Based at Iowa State University with domestic partners at Florida International University and the University of Maryland, CoSMIC is composed of an international consortium of universities and laboratories.

- **DANSE: The Data Analysis for Neutron Scattering Experiments** http://wiki.cacr.caltech.edu/danse/index.php/Main_Page Prompted by the development of the Spallation Neutron Source (http://www.sns.gov) (SNS), under construction in Oak Ridge, Tennessee, DANSE goals are to build a software system that 1) enables new and more sophisticated science to be performed with neutron scattering experiments, 2) makes the analysis of data easier for all scientists, and 3) provides a robust software infrastructure that can be maintained in the future.

- **DISUN: Data Intensive Scientific University Network** http://www.disun.org/ A grid-based facility of computational clusters and disk storage arrays distributed across 4 institutions each of which serves as a High Energy Physics Tier-2 site. The Compact Muon Solenoid (CMS) experiment is primarily supported by DISUN.

- **Ecoinformatics.org** http://www.ecoinformatics.org/ A community of ecologists and informatics specialists operating on the principles of open-source software communities. Collaborative projects such as software and standards development that cross institutional and funding boundaries are managed through contributed personnel in an open and democratic process. Ecological Metadata Language is a product of this community.

- **GriPhyN: Grid Physics Network** http://www.griphyn.org/ A large ITR project focused on computer science and grid research applied to the distributed computing and storage requirements of the high energy physics community. GriPhyN is a foundational element of the Open Science Grid (OSG).

- **ICPSRC: Inter - University Consortium for Political and Social Research** http://www.icpsr.umich.edu/ Maintains and provides access to a vast archive of social science data for research and instruction, and offers training in quantitative methods to facilitate effective data use. ICPSR preserves data, migrating them to new storage media as changes in technology warrant. ICPSR also provides user support to assist researchers in conducting their projects and identifying relevant data for analysis.

- **iVDGL: International Virtual Data Grid Laboratory** http://www.ivdgl.org/ A global data and compute grid serving physics and astronomy, funded as a large ITR project by MPS. While its sister project, GriPhyN, funded the grid research, iVDGL supports the operational grid and associated services supporting projects like CMS, ATLAS, and LIGO.

- **LHC: Large Hadron Collider and LCG: Large Hadron Collider Computing Grid** The Large Hadron Collider, http://lhc.web.cern.ch/lhc/ being built at CERN near Geneva, is the largest scientific instrument on the planet. The mission of the LHC Computing Project (LCG)

is to build and maintain a data storage and analysis infrastructure for the entire high-energy physics community that will use the LHC. See also the **World LHC Computing Grid** (WLCG) http://lcg.web.cern.ch/LCG/ , a distributed production environment for physics data processing.

- **LTER: Long-Term Ecological Research Network** http://lternet.edu/ A networked collaboratory of 26 field sites and network office focused on answering questions about the long-term dynamics of ecosystems ranging from near pristine to highly engineered sites. In addition to site-based science, education and outreach, five common research themes ensure that multi-site projects and synthesis are integral to the program.

- **NCEAS: National Center for Ecological Analysis and Synthesis** http://www.nceas.ucsb.edu/fmt/doc?/frames.html NCEAS and the National Evolutionary Synthesis Center (NESCent) http://www.nescent.org/main/ serve as data centers and collaboratories for the ecology and evolutionary biology communities. Both are involved in the development of software, house databases, and sponsor collaborative activities aimed as synthesizing research.

- **NEES: George E. Brown, Jr. Network for Earthquake Engineering Simulation** http://www.nees.org/ A shared national network of 15 experimental facilities, collaborative tools, a centralized data repository, and earthquake simulation software, all linked by the ultra-high-speed Internet2 connections of NEESgrid. These resources provide the means for advanced collaborative research based on experimentation and computational simulations of the ways buildings, bridges, utility systems, coastal regions, and geomaterials perform during seismic events.

- **NNIN: National Nanotechnology Infrastructure Network** http://www.nnin.org/ Integrated networked partnership of 13 user facilities that serves the resource needs of nanoscale science, engineering and technology. NNIN provides users with shared open access, on-site and remotely, to leading-edge tools, instrumentation, and capabilities for fabrication, synthesis, characterization, design, simulation and integration for the purpose of building structures, devices, and systems from atomic to complex large-scales.

- **OSG: Open Science Grid** http://www.opensciencegrid.org/ The national physics grid arising from the GriPhyN, iVDGL, and DOE's PPDG projects. As an operational grid spread across dozens of institutions and DOE sites, OSG primarily supports high energy physics grid applications and some non-physics applications in an opportunistic grid dominated by commodity clusters whose cpu counts total into the thousands. The OSG Consortium builds and operates the OSG, bringing resources and researchers from universities and national laboratories together and cooperating with other national and international infrastructures to give scientists from many fields access to shared resources worldwide.

- **QuarkNet Cosmic Ray eLAB Project** http://quarknet.uchicago.edu/elab/cosmic/home.jsp A distributed learning lab for collaborating high school physics students and teachers who collect and analyze cosmic ray data. The e-LAB participants work with computer scientists to provide cutting edge tools that use grid techniques to share data, graphs, and posters and to encourage collaboration among students nationwide.

- **SCEC CME : Southern California Earthquake Center Community Modeling Environment** http://epicenter.usc.edu/cmeportal/ A recent geophysics and IT collaboratory

targeted at seismic hazard analysis and geophysical modeling. The computational test-bed under development will allow users to assemble and run highly complex seismological and geophysical simulations utilizing Teragrid with the goal of forecasting earthquakes in Southern California.

- **UltraLight** http://ultralight.caltech.edu/web-site/ultralight/html/index.html A collaboration of experimental physicists and network engineers to enable petabyte-scale analysis of globally distributed data. Goals include: 1) developing network services that broaden existing Grid computing systems by promoting the network as a actively managed component; 2) testing UltraLight in Grid-based physics production and analysis systems; and 3) engineering a trans- and intercontinental optical network testbed, including high-speed data caches and computing clusters.

- **Veconlab: The Virginia Economics Laboratory** http://veconlab.econ.virginia.edu/admin.htm/ Focuses on game theory and social interactions in economics and related fields. The Veconlab server provides a set of about 40 web-based programs that can be used to run interactive, social science experiments for either teaching or research purposes.

- **Vlab: The Virtual Laboratory for Earth and Planetary Materials** http://www.vlab.msi.umn.edu/, An interdisciplinary consortium for development and promotion of the theory of planetary materials. Computational determination of geophysically important materials properties at extreme conditions provides accurate information to a) interpret seismic data in the context of likely geophysical processes and b) be used as input for more sophisticated and reliable modeling of planets.

Observatories:

- **CHESS: Cornell High Energy Synchrotron Source** http://www.chess.cornell.edu Provides users with state-of-the-art synchrotron radiation facilities for research in Physics, Chemistry, Biology, and Environmental and Materials Sciences. A special NIH Research Resource, called **MacCHESS,** http://www.macchess.cornell.edu/ supports special facilities for protein crystallographic studies.

- **CHRNS: Center for High Resolution Neutron Scattering** http://www.ncnr.nist.gov/programs/CHRNS/ Develops and operates state-of-the-art neutron scattering instrumentation with broad applications in materials research for use by the general scientific community. Combined, CHRNS instruments provide structural information on a length scale of 1 nm to ~10 microns, and dynamical information on energy scales from ~30 neV to ~100 meV, the widest ranges accessible at any neutron research center in North America.

- **EarthScope** http://www.earthscope.org/ Applies modern observational, analytical and telecommunications technologies to investigate the structure and evolution of the North American continent and the physical processes controlling earthquakes and volcanic eruptions. Efforts involve data transmission from numerous seismographs and GPS receivers to two data centers, serving the raw data in real time and making available derived data products via a web portal.

- **IceCube** http://www.icecube.wisc.edu/ A one-cubic-kilometer international high-energy neutrino observatory being built and installed in the clear deep ice below the South Pole Station. IceCube will open unexplored bands for astronomy, including the PeV (1015 eV) energy region, where the Universe is opaque to high energy gamma rays and where cosmic rays do not carry directional information because of their deflection by magnetic fields.

- **IRIS: Incorporated Research Institutions for Seismology** http://www.iris.edu/ A university research consortium dedicated to exploring the Earth's interior through the collection and distribution of seismographic data. IRIS programs contribute to scholarly research, education, earthquake hazard mitigation, and the verification of a Comprehensive Test Ban Treaty. The IRIS Data Management System manages and disseminates time series data from a variety of worldwide seismic instruments to provide basic data in support of earthquake studies and research into the structure of the Earth's crust, mantle and core.

- **LIGO: Laser Interferometer Gravitational Wave Observatory** http://www.ligo.caltech.edu Dedicated to the detection of cosmic gravitational waves and harnessing of these waves for scientific research. The facility consists of two widely separated installations within the United States — one in Hanford Washington and the other in Livingston, Louisiana — operated in unison as a single observatory.

- **Microbial Observatories** http://www.nsf.gov/bio/pubs/awards/mo.htm A network of sites or "microbial observatories" in different habitats to study and understand microbial diversity over time and across environmental gradients. Supported projects must establish or participate in an established, Internet-accessible knowledge network to disseminate information resulting from these activities.

- **NEON: National Ecological Observatory Network** http://www.neoninc.org/ A national research platform and virtual laboratory for studying the role of the biosphere in earths systems by examining the structure, function, and evolution of biological systems at regional to continental scales. A CI backbone will link field instrumentation and mobile instrument platforms, sensor networks, laboratory and field instrumentation, natural history archives, and analytical and modeling capabilities, to facilities for archival, computation, visualization, and forecasting.

- **NHMFL: National High Magnetic Field Laboratory** http://www.nhmfl.gov/ Develops and operates high magnetic field facilities that scientists use for research in physics, biology, bioengineering, chemistry, geochemistry, biochemistry, materials science, and engineering. High magnetic fields are a critical link in the development of new materials that impact nearly every modern technology.

- **ORION and OOI: Ocean Research Interactive Observatory Networks** http://orionprogram.org/ and http://www.orionocean.org/OOI/default.html An integrated observatory network for oceanographic research and education with interactive access to ocean observing systems that track a wide range of episodicity and temporal change phenomena. OOI has three elements: 1) a global-scale array of relocatable deep-sea buoys, 2) a regional-scaled cabled network consisting of interconnected sites on the seafloor, and 3) network of coastal observatories.

Other Virtual Organizations:

- **AToL: The Tree of Life initiative** http://www.phylo.org/AToL/ Supports the reconstruction of the evolutionary history of all organisms, seen as a grand challenge. The AToL research community works toward describing the evolutionary relationships of all 1.7 million described species.

- **CASA: Center for Adaptive Sampling of the Atmosphere** http://www.casa.umass.edu/ Integrates advancements in radar technology, networking and atmospheric sciences to develop a new, low cost network of weather radars that can adapt sampling procedures in real-time in order to optimize information and provide unprecedented details of severe storms.

- **CCMC: The Community Coordinated Modeling Center** http://ccmc.gsfc.nasa.gov/ Provides access to modern space science simulations and supports the transition to space weather operations of modern space research models. To support community research CCMC adopts state of the art space weather models that are developed by outside researchers, executes simulation runs with these models, offers a variety of visualization and output analysis tools, and provides access to coupled models and existing model frameworks.

- **CEDAR: Coupling, Energetics and Dynamics of Atmospheric Regions** http://cedarweb.hao.ucar.edu/cgi-bin/ion-p?page=cedarweb.ion For characterization and understanding of the atmosphere above ~60 km, with emphasis on the processes that determine the basic composition and structure of the atmosphere. Activities include collaborative research projects, multi-site field campaigns, workshops, and participation of graduate and undergraduate students.

- **CIG: Computational Infrastructure for Geodynamics** http://www.geodynamics.org/ Membership-governed organization that supports Earth science by developing and maintaining software for computational geophysics and related fields. CIG consists of: (a) a coordinated effort to develop reusable and open-source geodynamics software; (b) an infrastructure layer of software for assembling which state-of-the-art modeling; (c) extension of existing software frameworks to interlink multiple codes and data through a superstructure layer; (d) strategic partnerships with computational science and geoinformatics; and (e) specialized training and workshops.

- **CLIVAR: Climate Variability and Predictability** http://www.clivar.org/ An international research program addressing many issues of natural climate variability and anthropogenic climate change, as part of the wider (WCRP). It makes available a wide variety of climate-related data and models via a web interface and data management system.

- **CUAHSI HIS**: **The Consortium of Universities for the Advancement of Hydrologic Science, Inc.** http://gis.sdsc.edu/cuahsi/ Represents about 100 U.S. universities and develops infrastructure and services that support advancement of hydrologic science and education. The CUAHSI Hydrologic Information System (HIS) project is conducted by a group of academic hydrologists collaborating with the San Diego Supercomputer Center as a technology partner to produce a prototype Hydrologic Information System.

- **C-ZEN: Critical Zone Exploration Network** http://www.wssc.psu.edu/czen.htm Promotes interdisciplinary research on the zone defined by the outer limits of vegetation and the lower boundary of ground water – identified by the National Research Council as the Critical Zone (NRC, 2001). CZEN cyberinfrastructure consists of a critical zone ontology, development of tools for knowledge discovery and management, and data and metadata standards development.

- **DEISA: Distributed European Infrastructure for Supercomputing Applications** http://www.deisa.org/ A consortium of leading national supercomputing centers that currently deploys and operates a persistent, production quality, distributed supercomputing environment with European continental scope. DEISA's purpose is to enable discovery across a spectrum of science and technology through deep integration of existing national high-end computational platforms, coupled by a dedicated network and supported by innovative system and grid software.

- **EGEE: The Enabling Grids for E-sciencE** http://public.eu-egee.org/ Funded by the European Commission to build on recent advances in grid technology and develop a service grid infrastructure which is available to scientists 24 hours-a-day. Spanning over 30 countries and 150 sites across Europe, EGEE supports applications in Earth Sciences, High Energy Physics, Bioinformatics, Astrophysics, and other domains.

- **GCAT: Genomic Consortium for Active Teaching** http://www.bio.davidson.edu/projects/gcat/gcat.html A consortium of institutions whose purpose is to integrate genomic techniques in the undergraduate life sciences curricula. GCAT provides faculty with resources and training to utilize the technology and means to share data across institutions.

- **GEON: The Geosciences Network** http://www.geongrid.org/ Advances geoinformatics by training geoscience researchers, educators, students and practitioners in the use of cyberinfrastructure. GEON provides a service-oriented architecture (SOA) for support of "intelligent" search, semantic data integration, visualization of 4D scientific datasets, and access to high performance computing platforms for data analysis and model execution, via the GEON Portal.

- **GLOBEC: U.S. GLOBEC (GLOBal ocean ECosystems dynamics)** http://www.pml.ac.uk/globec/ Looks at how climate change and variability translate into changes in the structure and dynamics of marine ecosystems, marine animal populations and in fishery production. Research approaches use inter-related modeling, process-oriented studies, broad scale observations, and retrospective studies.

- **IPBIR: Integrated Primate Biomaterials and Information Resource** http://www.ipbir.org/ To assemble, characterize, and distribute high-quality DNA samples of known provenance with accompanying demographic, geographic, and behavioral information in order to stimulate and facilitate research in primate genetic diversity and evolution, comparative genomics, and population genetics.

- **LEAD: Linked Environments for Atmospheric Discovery**: http://lead.ou.edu/ Seeks to mitigate the impacts of extreme weather events by accommodating the real time, on-demand, and dynamically-adaptive needs of mesoscale weather research. A major

underpinning of LEAD is dynamic workflow orchestration and data management in a web services framework with grid-enabled systems.

- **MARGINS** http://www.margins.wustl.edu/Home.html Focuses on understanding the complex interplay of processes that govern creation and destruction of continental margins and result in the concentration of both resources and geohazards where land and oceans meet. Data from collaborative research projects are available via a Data Management System and web-based services.

- **NCAR: National Center for Atmospheric Research** http://www.ncar.ucar.edu/ A federally funded research and development center that, together with partners at universities and research centers, studies Earth's atmosphere and its interactions with the Sun, the oceans, the biosphere, and human society. NCAR hosts numerous projects that span the full range of VO activities and it recently launched a set of integrative initiatives designed to explore the Earth system from a 21st-century vantage point: http://www.ncar.ucar.edu/stratplan/initiatives.html

- **NCED: National Center for Earth-surface Dynamics** http://www.nced.umn.edu/ Fosters an integrated, predictive science of the skin of the Earth – "critical zone" - where interwoven physical, biological, geochemical, and anthropogenic processes shape Earth's surface. NCED's main cyberinfrastructure activities relate to a California field site at Angelo Coast Range Reserve with a steep, relatively rapidly eroding landscape where an advanced environmental observatory with a wireless network and automated environmental sensors are under construction.

- **NCN: Network for Computational Nanotechnology** http://www.ncn.purdue.edu/ Advances approaches and simulation tools that allow engineers to design new nanoelectronic and NEMS technologies. The fields include: i) nanoelectronics, ii) NEMS, and iii) nano-bioelectronics, with the overall goal of connecting dry electronic and mechanical nanosystems to wet biological nanosystems. **NanoHUB** http://www.nanohub.org/ is a web-based initiative spearheaded by NCN that provides online simulation tools for nanoelectronics and molecular dynamics, along with semiconductor devices, processes and circuits.

- **PEATNET: Peatland Ecosystem Analysis and Training Network** http://www.peatnet.siu.edu/ An international interdisciplinary collaborative effort focused on northern peatland ecosystems. Web site serves as a mechanism to coordinate PEATNET activities and communicate findings and outcomes, and as the location of a peatland Web-discussion group and the PEATNET Virtual Resource Center.

- **PRAGMA: The Pacific Rim Application and Grid Middleware Assembly** www.pragma-grid.net/mission.htm Formed to establish sustained collaborations and advance the use of grid technologies in applications among a community of investigators working with leading institutions around the Pacific Rim. In PRAGMA, applications are the key, integrating focus that brings together necessary infrastructure and middleware to advance the application's goals.

- **RIDGE** http://ocean-ridge.ldeo.columbia.edu/general/html/home.html A long-term research program to study Earth's oceanic spreading ridge system as an integrated whole, from its inception in the mantel to its manifestations in the biosphere and water column. Ridge 2000 supports interdisciplinary research and availability of data through web-based services.

- **SAHRA: The Center for Sustainability of semi-Arid Hydrology and Riparian Areas** http://www.sahra.arizona.edu/ Promotes sustainable management of semiarid and arid water resources, with three components: the SAHRA Geo-database (SGD) for data storage and sharing, hydrologic observations, and integrated modeling at three resolutions.

- **SBN: Seamount Biogeosciences Network** http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=0443337 Brings together the diverse science disciplines involved in seamount research for improved communication and scientific collaboration, data archiving and integration, and sharing of seagoing logistical operations. Key SBN components include regular workshops and the development of a community website and database.

- **Suominet: Real Time Integrated Atmospheric Water Vapor and TEC from GPS** http://www.suominet.ucar.edu/ Measures phase delays induced in GPS signals by the ionosphere and neutral atmosphere with high precision and converts these delays into integrated water vapor and total electron content (TEC). These measurements contribute to our understanding of the Earth's weather and climate system and TEC data addresses topics in upper atmospheric research.

- **TeraGrid** www.teragrid.org An open scientific discovery infrastructure combining leadership class resources at eight partner sites to create an integrated, persistent computational resource. Deployed in September 2004, TeraGrid brings over 40 teraflops of computing power and nearly 2 petabytes of rotating storage, and specialized data analysis and visualization resources into production, interconnected at 10-30 gigabits/second via a dedicated national network.

- **TESS: Time-sharing Experiments for the Social Sciences** http://www.experimentcentral.org/ Permits original data collection through national telephone surveys to which researchers can add their original questions and through arrangements that allow researchers to run their own studies on random samples of the population that are interviewed via the Internet.

- **UNAVCO** http://www.unavco.org/ A non-profit, membership-governed consortium that supports and promotes high-precision techniques for the measurement and understanding of crustal deformation. The data processing center analyzes and distributes GPS data from continuous GPS installations globally, including the dense array of stations dedicated to the EarthScope project.